Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Final Report

Project Number: 64

Emotional Talking Face Generation System using DNN

Author: Gayeon Kim

Project Partner: Yugyeong Hong

Supervisor(s): Ho Seok Ahn, Trevor Gee

13/10/2023

Declaration of Originality

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

Name: Gayeon Kim

ABSTRACT

The primary goal of our project is to generate artificial humanoid avatars, specifically talking faces from a single image and text for enhanced human-robotic interaction. We put a specific emphasis on avatars that exhibit precise lip motion, head movement, and dynamic facial expressions. We believe that these attributes are essential components, making avatars significantly more engaging to human users. Contrary to traditional 3D modelling techniques that are commonly used in many modern state-of-the-art systems, our project aims to build avatars from machine-learned image augmentations. While numerous studies have been conducted on talking face generation systems, most have explored lip motion in isolation from emotional facial shifts. Additionally, a significant dependency on audio or video inputs can be seen as a limiting factor in many methods. This work will explore the leading techniques for generating realistic emotional talking faces. We introduce EmoFaceGen, a step-by-step approach that distinctly processes emotion and lip synchronisation stages for enhanced accuracy and effectiveness. Both qualitative and quantitative evaluations show that our system outperforms other open-source models. While there is room for further improvement, we view our model as a significant advancement, especially when considering the memory and hardware limitations associated with conventional 3D graphics methods.

Acknowledgements

I would like to thank my project supervisors, Ho Seok Ahn and Trevor Gee for their guidance and support during this research project.

Table of Contents

1. Introduction

In the contemporary digital era, the generation of talking faces has become a topic of significant interest due to advancements in technology that have made it feasible to create interactive, realistic faces. These developments are about more than just visual fidelity involving the capabilities of artificial intelligence to create digital avatars that can simulate human-like interactions, a task that was not achievable with past technologies. The importance of this research domain is further emphasised by the projected growth of the global Human-Machine Interface (HMI) market, which is expected to escalate from USD 4.9 billion in 2022 to USD 7.3 billion by 2027 at a compound annual growth rate (CAGR) of 8.1% during the forecast period [1]. This statistic not only highlights the increasing economic significance of human-machine interactions but also underscores the necessity for more natural and relatable communication with devices, moving beyond the traditional methods that often involve navigating through dense manuals or learning complex procedures.

Talking face generation plays an essential role in this aspect, offering a human-like channel for interaction that is both familiar and intuitive, significantly reducing the unease and learning difficulties often associated with innovative technologies. As a result, researchers have conducted numerous experiments and explored various techniques to create talking faces. With the rapid development of AI and the emergence of technologies such as deepfake, the significance of generating talking faces cannot be understated. Talking face generation not only creates realistic visual content but also improves human-computer interaction, benefits medical scenarios, and provides diverse entertainment experiences.

Moreover, with the growing popularity of virtual reality and related platforms, the demand for high-quality talking face generation is expected to rise, emphasising the importance of this research

domain. However, despite these growing interests and extensive research, there remain several challenges in the realistic talking face generation. These include the system's capability to adjust emotions, implement head pose movement, utilise just one facial image and text-based input, and function with low hardware cost.

Most recent studies have explored lip motion in isolation from emotional facial shifts and head movements. Emotions play significant roles in communication by altering the nuances and reception of messages. [2] emphasises the crucial role of emotional involvement in improving the clarity of communication, illustrating that errors in human-robot interactions exceed 30% when emotional signals are absent. Consequently, as critical elements of conversation, it is essential to render talking faces with a range of emotions, such as happiness, sadness, and anger. Head movement is also vital as it adds naturalism to generated videos, making interactions more genuine. Furthermore, many current techniques depend on audio or video inputs, a requirement that presents limitations. This is not only because some individuals may not have access to recorded audio or filmed videos but also because videos, in particular, are large files that consume significant storage space and are slow to transmit. Therefore, leveraging just a single image combined with text input appears more universally accessible.

Recognising these needs, we propose an innovative system that utilises a target face image and text dialogue to generate a realistic emotional talking face video. Our research focuses on developing realistic talking faces, offering a unique opportunity for individuals to converse with representations of lost loved ones. While this goal is not an immediate outcome of our current work, it is a key objective we aim to achieve in our future plans. Additionally, we aim to enhance human-robot interaction, improving communication capabilities in social robots. The primary focus of our project is to achieve realistic visual quality in talking faces, ensuring that messages are conveyed accurately

during communication. Our system offers the flexibility to control emotional tones and head movements. By incorporating a text input figure into our system, we have also improved its accessibility, significantly expanding its potential applications. This report will concentrate on the system's lip synchronisation, head movement and visual quality improvement as my project partner focused on emotion generation.

In the subsequent sections of this paper, we provide a structured presentation of our research journey and experimental findings. Section 2 offers a review of related works, setting the foundation and context for our study. In section 3, we outline our research methods, beginning with our initial End-to-End approach and its methodology, followed by our refined Step-by-Step approach, which represents our final method. We provide a thorough discussion of the Step-by-Step methods, including their steps and the criteria used for evaluation. Moving to Section 4, we detail the experiments conducted and present their results. Section 5 offers a discussion of the findings detailed in Section 4, with an emphasis on the advantages and limitations observed. Finally, in Section 6, we conclude the paper by summarising our discoveries and suggesting potential areas for further improvements.

2. Related works

Our research integrates several areas of image augmentation, including facial expression changes, lip movements, head position dynamics, and visual quality improvement. In this section, key literature for lip synchronisation, head pose implementation and visual quality improvement that supports our approach is outlined.

2.1. Lip synchronisation

When it comes to generating a realistic talking face, lip synchronisation is an extremely important component. It is necessary to synchronise the motions of the lips with the sounds that are being produced by the voice to give the impression that one is speaking. Moreover, this precise synchronisation supports comprehension, considering that people can interpret speech from lip movements alone, even in the absence of sound, highlighting the impact of visual cues in speech perception. Extensive research has been undertaken in the field of lip synchronisation, resulting in the proposal of various techniques. Some of the research papers propose methods using a generative adversarial network (GAN). GAN consists of two neural networks: the generator, which generates a single image, and the discriminator, which distinguishes the image produced by the generator from real data. GAN has the advantage of being able to generate new samples that are highly realistic, as they use stochastic noise as input instead of relying on expected data distribution [3]. Consequently, this can ensure more accurate and smooth lip movements in sync with the given audio. [4] proposed a model called LipGAN, which uses a GAN architecture that utilises both audio and visual features to generate lip-synced video frames. [4] used consecutive frames from the input video and its corresponding audio input, which was converted into melspectograms to predict the lip shape for the next frame. However, the use of pixel-wise loss function in this model resulted in generalising predictions over truly learning the correlation of lip shapes to sound. To improve upon this finding, [5] introduced a model called Wav2Lip, which is a widely recognised lip synchronisation technique. Initially, Wav2Lip focused on its discriminative training strategy, which is employed to ensure the appropriate mouth shape by determining if the lips in the video were in sync with the provided audio. However, it still comes with numerous shortcomings. The Wav2Lip model generates videos at a relatively low resolution of 96 × 96 pixels, causing the mouth region to appear

blurry and lack finer details. Then, [5] proposed to integrate GAN architecture into the Wav2lip model. The use of the visual quality discriminator model of GAN showed a slight improvement in lip synchronisation but did not meet the expected outcomes.

[6] proposes an approach called a Sadtalker that implements Wav2Lip during the initial phase, demonstrating superior performance in lip synchronisation. This model uses Wav2Lip to generate lip motions corresponding to sound. Then 3D Morphable Model (3DMM) technique is utilised to reconstruct 3D face structure, allowing for the capture and rendering of various facial expressions, including lip movements.

2.2. Head pose implementation

The movement of the head is also an important factor for generating a realistic talking face since the face may appear unnatural without the head movement. While Wav2lip [5] and other lip synchronisation methods lack the ability to generate head movement on their own, we plan to incorporate head movement techniques into our project.

2.2.1. Visual vs Audio based

There are two approaches for generating head movement in talking face generation: one involves using a pose source video, while the other involves learning pose motions directly from audio input. [7] introduces a deep learning-based architecture to predict facial landmarks, capturing overall head poses from only speech signals. Meanwhile, studies such as [8][9] leverage pose source videos specifically to adjust for head motions. [9], in particular, notes the difficulty of inferring head pose from audio signals alone. As a solution, they propose incorporating the pose sequence estimated from a training video clip as an additional input to the model.

2.2.2. 2D and 3D based

For incorporating the pose from a source video, the early works, including studies like FOMM [10] and DVP [11], usually focused on 2D-based methods or 3D-based methods which extract 3D Morphable models (3DMM). However, they are subject-dependent and cannot be generalised, which leads to restricting their applicability in varied real-world scenarios.

2.2.3. Disentanglement of pose and expression

[12] proposes a model called the Disentanglement of Pose and Expression (DPE), which is a self-supervised disentanglement framework that does not rely on paired data or predefined 3DMMs. A component called the Motion Editing module focuses on the disentanglement of pose and expression in the latent space using multiple perceptron layers. Given a source image, a driving image, and an editing indicator, this module produces an edited latent code with multiscale feature maps of the source image. By not relying on predefined 3DMM, this model offers more flexibility and generalises better to diverse faces.

2.3. Visual quality improvement

A Codebook lookup transformer (Codeformer) [13] proposes a methodology that enhances face restoration from degraded inputs through a three-tiered approach. Initially, an autoencoder is trained using vector quantisation, forming a discrete codebook that maps low-quality inputs to high-quality outputs. Then, a Transformer module and a Controllable feature transformation module are incorporated to predict the corrupted or missing parts and optimise the results across varying levels of degradation.

3. Research methods

3.1. System Approach

Given a face exemplar, a target face, an emotion condition, and a dialogue in a text form, two pipelines are developed. Text inputs are processed using Google's text-to-speech API [14]. There are two pipelines developed for the generation processes. Upon recognising the limitations of the first pipeline, we shifted our focus to the second one.

3.1.1. End-to-end approach (first pipeline)

Initially, our method was designed to adjust emotions, synchronise lips, and implement head movements using a single model. The comprehensive flow of this approach is illustrated in Fig. 1. We built our approach on the foundation of a model called EmoTalkingFace, proposed in [15], which was already equipped for emotional face generation. However, it fell short in realistic visual quality compared to other face generation models. Coupled with its prolonged training duration, it made us reconsider its efficiency, leading us to employ a more sequential strategy.

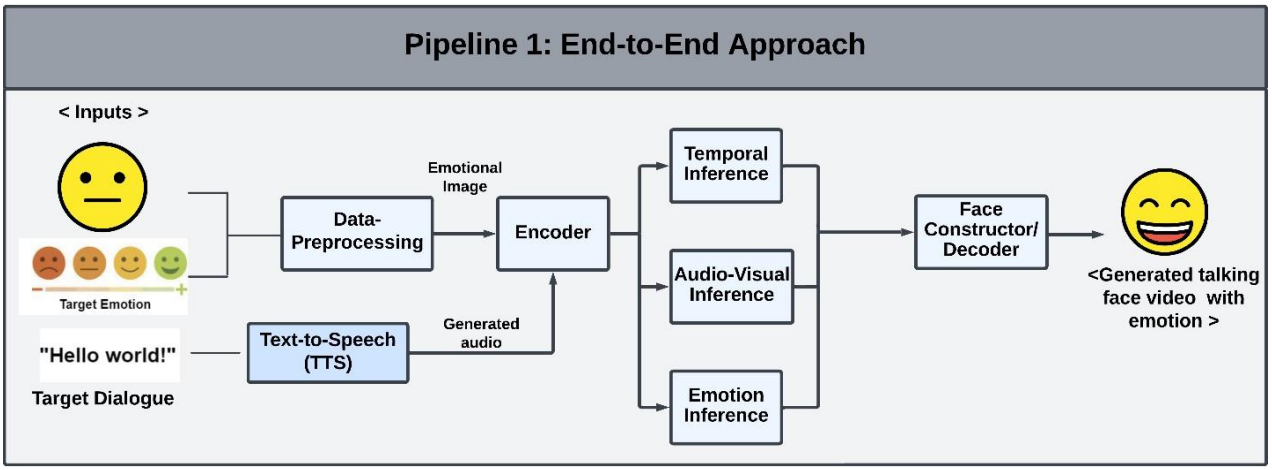

Fig. 1. First pipeline

3.1.2. Step-by-step approach (second pipeline)

The updated pipeline segregates emotion generation, lip synchronisation and head movement implementation across five distinct phases, each powered by its specific model. The following discussion provides a detailed explanation of each of these phases. This progression is depicted in Fig. 2.

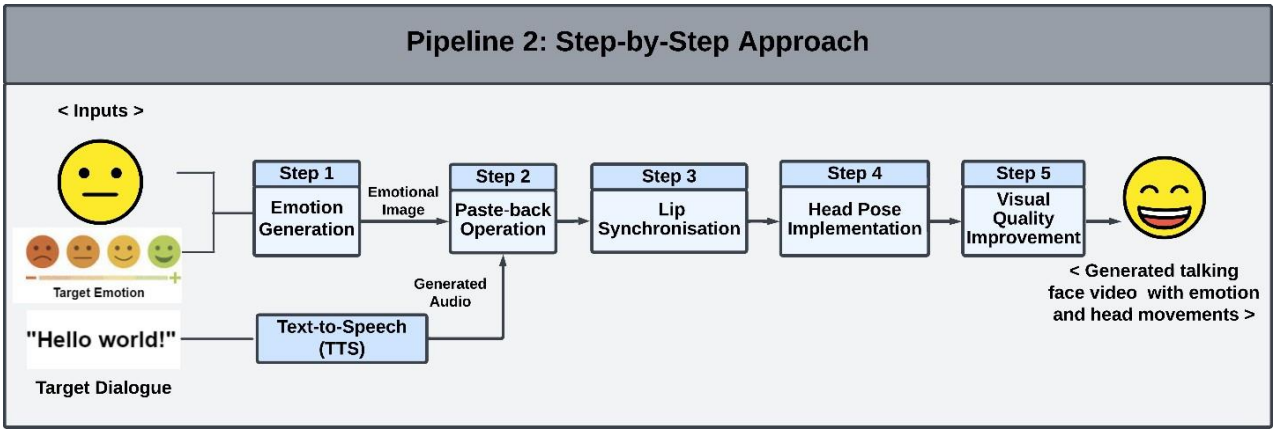

Fig. 2. Second pipeline

The first step in Fig. 2, which focuses on emotion generation, was completed by my project partner.

Step 2: Paste-back operation

Implementing head pose movement in step 4 into our lip-synchronised video introduced an issue related to restoring the image to its uncropped state. This is because the images processed during the emotion stage were specifically cropped to focus solely on the face. We assumed that this is because the head pose implementation stage captures the entirety of the head movement. Consequently, the cropped face lacks essential components and spatial references around the face, such as the neck and shoulders. To address this, before proceeding to the head pose stage, the past back operation needed to be performed. We refer to this as the paste-back operation, where the processed face with emotion is merged back into its original image to ensure accurate head movement.

The paste-back operation is implemented prior to lip syncrhonisation process because it must be applied at the image level, whereas the lip synchronisation stage involves video processing. To perform the paste-back operation, facial landmarks for both images are identified using the dlib library's frontal face detector and shape predictor. By utilising the angles between the eyes, we achieve face alignment to maintain a consistent horizontal orientation. Then, the facial region from the cropped image is extracted and overlaid onto the corresponding region in the original image. To guarantee a flawless fusion between the inserted face and the background of the original image, we employ soft masks on both faces. The mask undergoes Gaussian blur to ensure smooth transitions between the face and the background during the cloning process. OpenCV's 'seamlessClone' function is used to blend the face from the cropped image into the corresponding region in the original image.

Step 3: Talking Face Video Generation with Lip Synchronization

In the lip synchronisation phase of our project, we leverage pre-trained weights and models from Sadtalker [6]. Within Sadtalker, the exclusive lip motion coefficients are used as the target, working through the previously established network of Wav2Lip [5]. A 3DMM method reconstructs 3D facial structure, emphasising lip movements. This approach allows us to achieve accurate lip synchronisation. Sadtalker was originally designed for rendering various facial expressions, including details such as eye blinks and eyebrow movements. For our project, we selectively extract only the lip, ignoring other facial expressions to concentrate solely on lip synchronisation.

Step 4: Head pose implementation

In the head pose implementation phase of our project, we currently leverage DPE [12]. We employ a pre-trained model that has been trained on the voxCeleb dataset [16], which includes over 100K videos of 1,251 subjects.

Step 5: Visual quality improvement

The lip synchronisation phase in our project results in reduced overall image quality. Even though Sadtalker [6] already employs Adversarial Loss, Reconstruction Loss, Perceptual Loss, Expression Loss, and to produce high-quality images that are not only similar in pixel values to the original but also in terms of perceptual quality and specific features, the produced images still have some visible noises and imperfections. Instead of introducing an additional loss function that could potentially make the system more complex, we have integrated the Codeformer [13] to enhance the visual quality.

3.2. Evaluation

To assess the realism of integrating lip synchronisation and head movement, we undertook a qualitative evaluation focusing on the naturalness of the generated videos. When evaluating the performance of lip synchronisation, both qualitative and quantitative evaluations were conducted. For the qualitative evaluation, they were shown a series of unlabelled videos that are produced by our system as well as other models.

For the quantitative evaluation, we employed various metrics that evaluate different aspects of video quality, including elements like noise ratios and structural similarity, in comparison to real sources.

- MRE (Mean Relative Error): This metric quantifies the average error between the generated video and a real source in terms of a percentage, highlighting how far the generated output is from the real data. A lower MRE value indicates the generated data is a closer match to the real source. However, it might not be able to capture precise structural differences.

- PSNR (Peak Signal-to-Noise Ratio): This metric is used to assess the quality of reconstructed images or videos. PSNR value is measured in decibels (dB), and a higher value indicates better the quality of the reconstructed output. While a high PSNR indicates less distortion, it does not always correlate with human evaluations of visual quality.
- SSIM (Structural Similarity Index Measure): This metric measures the visual impact of errors in the structural information of the video. The SSIM value ranges from -1 to 1, where one indicates that the two sources being compared are identical. While SSIM can capture structural changes effectively, it might not be sensitive to some textural differences.
- FID (Frechet Inception Distance): This metric measures the distance between the data distribution of the generated output and the real source in the feature space. A lower FID score indicates that the two sources are more similar. Although FID is a popular metric in the deep learning community, especially for tasks like measuring similarity between two sets of images or videos, it depends on specific model architectures.

4. Experiments and results

4.1. Software and hardware setups

The development process was undertaken utilising resources provided by the university, which includes a designated university lab computer and Docker environment equipped with a Quadro P6000 GPU and 24576MiB of RAM. Additionally, Google Colab Pro was used, featuring a T4 GPU and 32GB of RAM.

4.2. Lip synchronization

Fig. 3 below presents the outcomes of our lip synchronisation process with two different audio inputs: ‘this’ and ‘four’. The visual representation clearly shows that the lip motion matches the

corresponding audio input. This synchronisation ensures that the visual movements of the lips match up smoothly with the sound, offering a consistent experience for viewers.

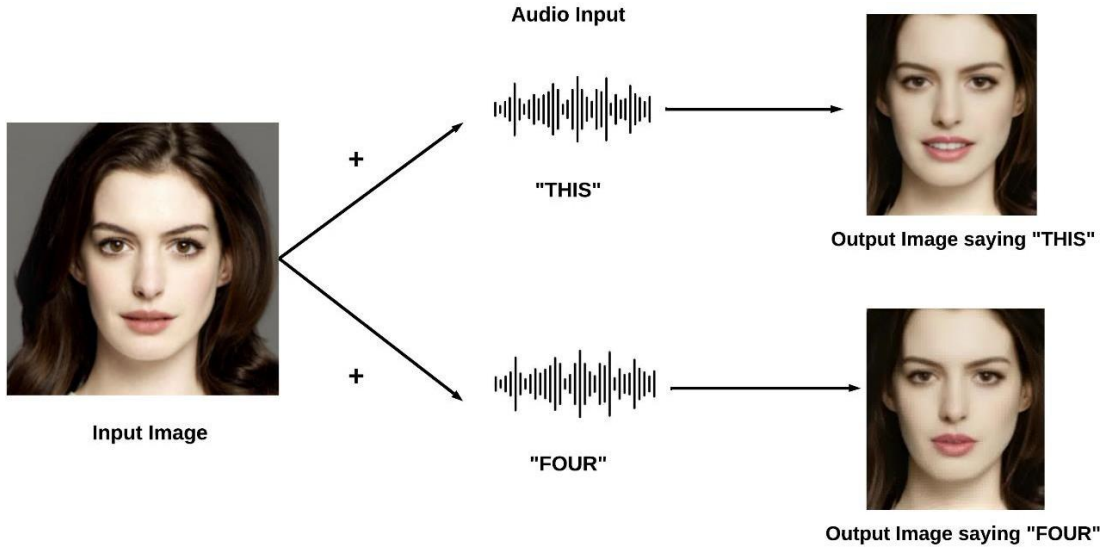

Fig. 3. Lip synchronisation result

Instead of solely relying on visual assessments, we employed both quantitative and qualitative evaluations. For the quantitative evaluation, we referred to the metrics described in section 3.2. Two other state-of-the-art methods, PC-AVS [8] and Video-retalking [18], are included in this assessment to compare our approach with the latest advancements. The evaluation results can be seen in Table 1 below.

Table 1: lip synchronisation quantitative evaluation.

	PSNR ↑	MSE ↓	SSIM ↑	FID ↓
SPT Wav2Lip	33.791	82.680	0.945	19.609
PC-AVS	14.985	6197.761	0.453	222.123
Video-retalking	**34.958**	62.498	0.955	74.156
Sadtalker+ Codeformer	32.245	**53.894**	**0.967**	**18.042**

Our approach achieved top scores with an MSE of 53.894, an SSIM of 0.967, and an FID of 18.042. However, when it came to PSNR, it ranked third, scoring roughly 2.7 points below the video-retalking method.

For the qualitative evaluation, we carried out a survey among 10 participants aged between 18 and 50. This age range was selected to ensure diverse feedback, capturing different perspectives and experiences related to lip synchronisation quality. A scale ranging from 0 to 10 is used, where 0 indicates very poor lip synchronisation quality, and 10 signifies the perfect match. The videos presented in the survey included real sources, results from the SPT Wav2Lip [17] method, and outputs from our system. SPT Wav2Lip is an improved version of the Wav2Lip [5], developed by students from the previous year. In the bar graph presented in Fig. 4, videos from real sources are labelled as ‘Ground truth’, results from SPT Wav2lip as ‘SPT Wav2lip’, and outputs from our system as ‘EmoFaceGen’. The average survey score for the ground truth video was 9.1. While the SPT Wav2lip method scored 5.7, our method received 6.7.

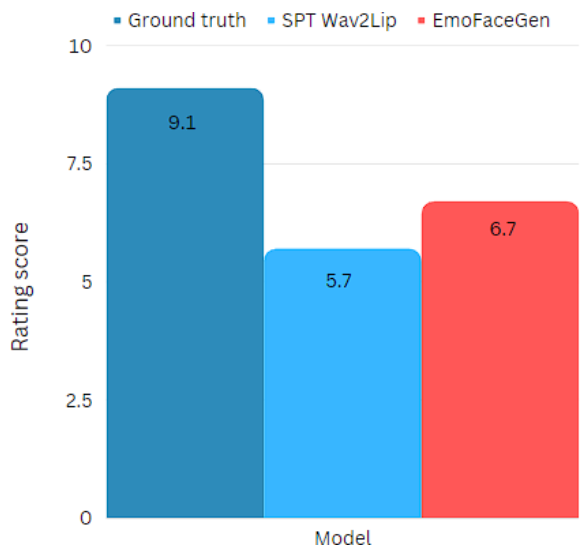

Fig. 4. Lip synchronisation quality survey results.

4.3. Paste-back operation

In our initial attempt at the paste-back operation, we focused on cropping the central facial region and overlaying it onto the original image. To achieve this, the Haarcascades face detection method was used to identify and extract the face regions from both images, making adjustments to exclude the forehead and chin. Although Gaussian blur was applied to smoothen the boundaries, the result lacked naturalness, with clear edges noticeable in Fig. 5(a). The difference in size between the cropped face and the original face added challenges.

For our second attempt, we focused solely on key facial features such as eyebrows, eyes, nose and mouth, given their significance in expressing emotion. While the same face detection and landmark initialisation were used in our final approach, only key facial features were extracted. However, this approach also faced challenges. Although the mouth region aligned appropriately, the placement of the eyes, eyebrows and nose was not consistent, as depicted in Fig. 5(b). The inconsistency could have arisen due to positional and scale differences between the two faces.

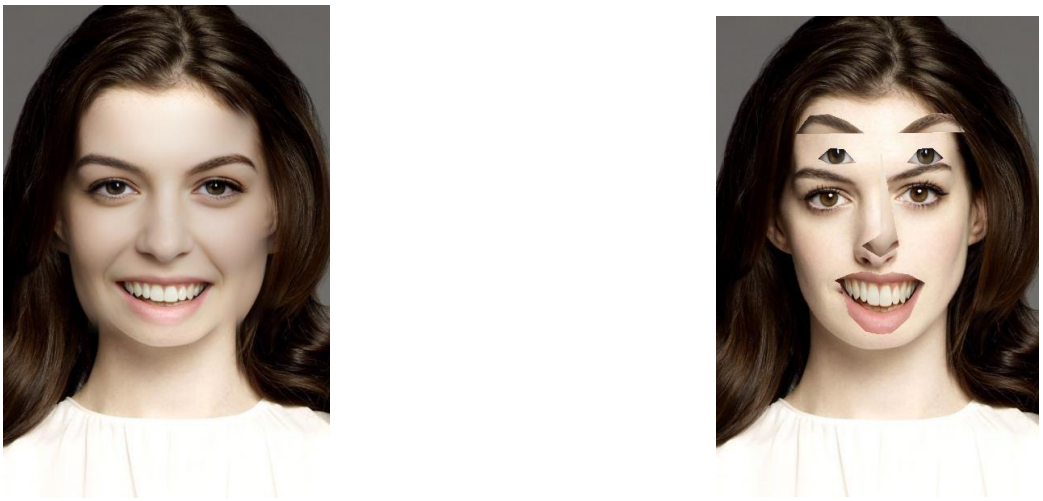

Fig. 5. (a) initial approach of pasting the cropped face to the original image, (b) second approach of paste-back operation.

The final approach, as detailed in Section 3.1.2 Step 3, successfully placed the cropped face onto the original image. The outcome of our final method is in Fig. 6.

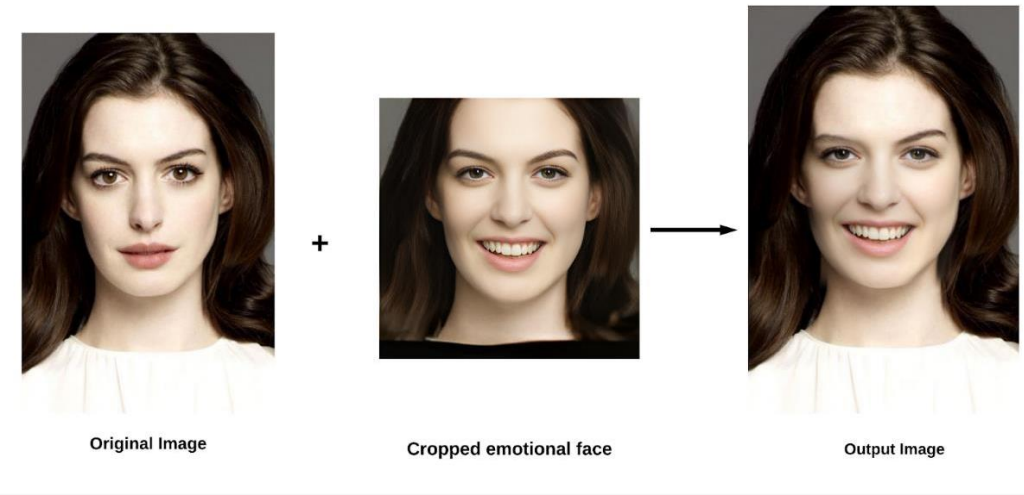

Fig. 6. Final paste-back operation results

4.4. Head pose implementation

Our initial approach of combining the lip synchronisation technique and the head pose implementation was to set the head pose prior to the lip synchronisation phase. However, it resulted in not recognising the mouth well, and the lip synchronisation process struggled to accurately match the mouth movements to the audio source. We theorised that this is because the head pose adjustments caused spatial shifts that disrupted the alignment and positioning of the mouth. Given these results, to retain the precision of lip movements, we changed our methodology. The head pose implementation is placed as a post-procedure, prioritising lip synchronisation before making head pose adjustments to maintain accurate lip movements.

Fig. 7. (a) initial approach of implementing head pose first (the mouth never appears open), (b) final approach of implementing head pose as a post-processing procedure.

Fig. 7(a) presents the outcome of our initial approach wherein the head pose is applied first, leading to the challenge of the mouth moving without opening. Conversely, Fig. 7(b) displays the results of our refined methodology, which shows correct and desirable outcomes.

4.5. Overall system from steps 2 to 5

A comprehensive qualitative evaluation, including steps 2 through 5, was carried out to assess the integration of lip synchronisation, head pose movement and visual quality improvement. Similar to lip synchronisation, we conducted a survey where participants were asked to rate videos from both real sources and those generated by our system, using a scale ranging from 0 to 10. Our method was evaluated for realism against the ground truth video in the overall system assessment. The ground truth video received an average score of 9.1, and our method achieved a rating of 6.0, as shown in Fig. 8.

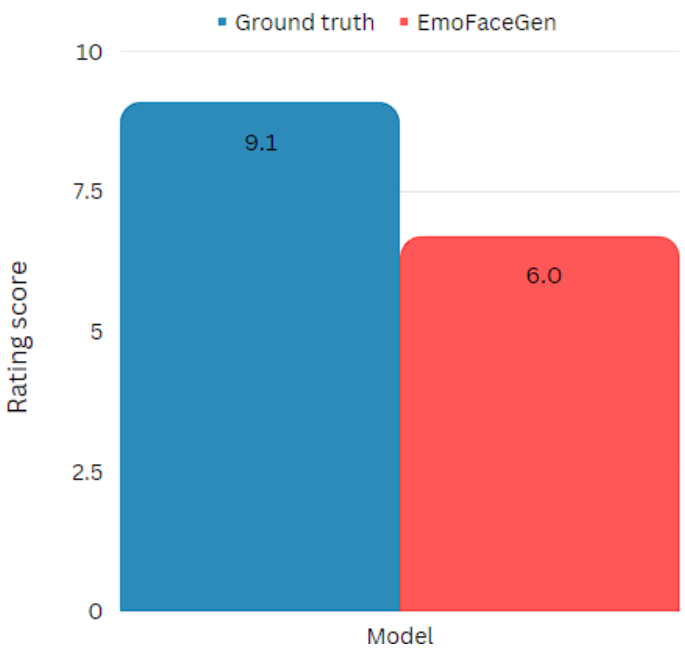

Fig. 8. Survey results for a combination of lip synchronisation and post-processing procedures quality.

5. Discussion

5.1. Lip synchronisation

By leveraging the lip synchronisation component from Sadtalker, we were able to encode an input image and an audio source for accurate alignment of lip movements. The result for lip synchronisation can be seen in Table I and Fig. 4. Table I shows that across three out of four metrics, our approach performed the best, outperforming the other state-of-the-art techniques in the evaluation. While the PSNR value for our method was lower than one of the other methods, this might be caused by the enhanced quality of the generated video due to the use of Codeformer [13]. A limitation of PSNR is that it does not always align well with human visual perception, as mentioned earlier. Therefore, a lower PSNR does not necessarily imply worse visual quality, especially when structural improvements are made in the generated output.

Furthermore, survey results in Fig. 4 show a clear difference between real videos and generated ones. However, they emphasise the improvements our approach has achieved, outperforming the SPT Wav2Lip [5] model- an improved version of the original Wav2lip. Although certain scores indicate our method could benefit from further refinement, a significant number of responses highlighted the accuracy of mouth shapes in our videos, leading to high scores from many participants.

Even though we successfully achieved lip synchronisation, the overall image quality was reduced, leading to a less distinct mouth area. Furthermore, the lack of head movement resulted in an unnatural appearance in the talking face. As a result, the next steps involved incorporating head movement and enhancing visual quality.

5.2.Overall system from steps 2 to 5

The results in Fig. 8 show a clear difference between the real videos and those generated by our system, with the latter receiving marginally lower scores. A potential reason for this could be the

certain unnatural appearances caused by the head pose integration and the efforts to improve visual quality. However, it is worth noting that the generated videos still achieved a score of 6.0, indicating a performance leaning more towards the positive end of the scale.

There are some possible refinements that could be made to improve the performance of our system. Our system currently has a limitation where it only effectively recognises and processes faces that are oriented straight on, both in terms of emotion generation and head pose implementation. Implementing a head pose on non-frontal faces is feasible, but it slightly modifies the facial appearance. To enable our system to work on non-straight faces, we propose integrating alignment tools as a pre-processing step. By repositioning the faces to a more neutral orientation, we aim to not only increase the accuracy but also minimise any unintended alterations to facial appearance.

6. Conclusion

Our project aimed to generate a realistic talking face generation from a single image and an audio source converted from text. Our initial approach was to design a single model that could generate emotions and synchronise lip movements with the audio. However, this approach resulted in unrealistic visual quality and extended the training duration unnecessarily. After recognising these limitations, we decided to change to the second approach, which enabled us to generate more realistic results in terms of visual quality and lip synchronisation.

In our second approach, emotion-rendered image was obtained first. This image was advanced to the lip synchronisation stage, where lip movements were generated to correspond with the audio. Despite the successful lip synchronisation, the absence of head movements, which add realism to the talking face, was noticeable. To address this, the implementation of head movements emerged as an essential next step. Furthermore, due to noticeable decreased visual quality in the initial stages, we committed

to visual quality enhancement. Initially, we attempted to incorporate head movement before lip synchronisation. This approach proved ineffective because the system struggled to accurately detect the mouth, leading to animations where the mouth moved without opening. Consequently, we adjusted out method to add head movements after obtaining lip synchronisation.

A significant challenge encountered was the necessity for a paste-back operation prior to the head movement implementation. As the emotion generator was designed to crop only the face, it limited the head pose module's accuracy due to insufficient surrounding contextual information. To address this issue, the cropped face was repositioned onto the original image before executing the head pose procedure. For visual quality enhancement, a transformer model was utilised. This model was designed to identify and enhance lower-quality aspects of the image, drawing on the data from the higher-quality regions. This methodology allowed for a more refined and realistic final result.

Both quantitative and qualitative evaluations revealed that our method outperformed other state-of-the-art methods and the work done last year. Furthermore, the overall quality of our system has been evaluated to be reasonably high. We concluded that our system is appropriate for use cases such as human-computer interaction or talking to departed loved ones. While our system provided promising results, our future work aims to expand its applicability to non-frontal facial orientations using alignment tools to minimise alterations. Furthermore, we aim to further enhance the realism of the generated faces, making them more indistinguishable from actual human representations. Another key area of future work involves incorporating head motion that responds dynamically to different emotions, adding a significant aspect of realism and expressiveness to the interactions.

References

[1] MarketsAndMarkets, “Human Machine Interface Market by Product (Hardware (Basic HMI, Advanced PC Based HMI, Advanced Panel Based HMI) and Software (On Premise HMI and Cloud Based HMI), Configuration (Embedded HMI, Standalone HMI) and Region - Global Forecast to 2027”, Aug. 2022, https://www.marketsandmarkets.com/Market-Reports/human-machine-interface-technology-market-461.html

[2] Yang, Yiju & Williams, Andrew, “Improving Human-Robot Collaboration Efficiency and Robustness through Non-Verbal Emotional Communication”, Mar. 2021, doi: 10.1145/3434074.3447191.

[3] Cheng Jieren, Yang Yue, Tang Xiangyan, Xiong Naixue, Zhang Yan, Lei Feifei, “Generative Adversarial Networks: A literature review”, 2020, KSII Transactions on Internet and Information Systems, Volume 14 Issue 12, Pages. 4625-4647

[4] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, C. V. Jawahar, “Towards Automatic Face-to-Face Translation”, Mar. 2020, doi: https://doi.org/10.1145/3343031.3351066

[5] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C V Jawahar, “A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild”, Aug. 2020, doi: https://doi.org/10.48550/arXiv.2008.10010

[6] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, Fei Wang, “SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation”, 2023

[7] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, Dingzeyu Li, “Make it talk: Speaker-Aware Talking head animation”, Feb. 2021, doi: https://doi.org/10.1145/3414685.3417774

[8] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, Ziwei Liu, “Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation”, Apr. 2021, doi: https://doi.org/10.48550/arXiv.2104.11116

[9] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, Xun Cao, “EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model”, Sep. 2022, doi: https://doi.org/10.48550/arXiv.2205.15278

[10] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, Sergey Tulyakov, “Motion Representations for Articulated Animation”, Apr. 2021, doi: https://doi.org/10.48550/arXiv.2104.11280

[11] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu, “Photorealistic audio-driven video portraits”, Dec. 2020, doi: https://doi.org/10.1109/TVCG.2020.3023573

[12] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, Dong-ming Yan, “DPE: Disentanglement of Pose and Expression for General Video Portrait Editing”, Mar. 2023, doi: https://doi.org/10.48550/arXiv.2301.06281

[13] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, Chen Change Loy, “Towards Robust Blind Face Restoration with Codebook Lookup Transformer”, Nov. 2022, doi: https://doi.org/10.48550/arXiv.2206.11253

[14] Google Cloud, “Cloud Text-to-speech.” Google., 2017.

[15] S. E. Eskimez, Y. Zhang and Z. Duan, "Speech Driven Talking Face Generation From a Single Image and an Emotion Condition," in IEEE Transactions on Multimedia, vol. 24, pp. 3480-3490, 2022, doi: 10.1109/TMM.2021.3099900.

[16] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, “Voxceleb: a large-scale speaker identification dataset”, 2017, doi: https://doi.org/10.21437/Interspeech.2017-950

[17] Takahiro Ishiguro, Alan Lin, Jong Yoon Lim, Trevor Gee, Edmond Liu, Bruce A. MacDonald, and Ho Seok Ahn, “Talking Face Generation from Facial Image, Target Emotion and Speech Dialogue”, In Proceedings of the 2022 Australasian Conference on Robotics and Automation (ACRA 2022), Dec. 2022.

[18] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, Nannan Wang, “VideoReTalking: Audio-based Lip Synchronization for Talking Head Video Editing In the Wild”, Nov. 2022, doi: https://doi.org/10.48550/arXiv.2211.14758