Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Literature Review and

Statement of Research Intent

Project Number: 64

Talking face generation system using DNN

Gayeon Kim

Yugyeong Hong

Hoseok Ahn, Trevor Gee

19/04/2023

Declaration of Originality

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

Gayeon Kim

Name: Gayeon Kim

ABSTRACT: The literature review explores existing techniques as well as the challenges that exist in the process of generating talking faces with emotional expressions. Numerous studies have been conducted on talking face generation systems, but there has been a lack of research on talking face generation with emotions and there has been no research that has resulted in a perfect output that meets our expectations. We will be researching methods for the system that takes inputs such as images, video, text, or audio to generate a talking face with various emotions. The review provides a comprehensive overview of the current state of talking face generation with emotion and identifies important topics for future development.

1. Introduction

In recent years, talking face generation has become a topic of significant interest, and numerous experiments have been conducted to investigate various strategies for generating talking faces from a single image. However, there seems to be a lack of research on generating faces with emotions. Since emotion plays a significant role in human communication and is therefore one of the most important conversational elements, we intend to generate talking faces with a range of emotions, such as happiness, sadness, and anger. We will continue the research conducted by other students last year for our project. By analysing last year's project completed by other students, we have identified significant areas for improvement and additional features that could be added to generate realistic talking faces. Based on my research, I have identified five key features that must be considered to develop a realistic talking face generation system. The project objectives section below will provide a detailed understanding of the project goal and its five key features.

2. Project objectives

Our project aims to generate a realistic talking face with emotions, using a single image and text as inputs. The following are the goals (five key features) we want to achieve in this project.

- Achieving high quality and natural lip synchronisation
- Incorporating various ranges of emotions into talking faces such as happiness, sadness, fear, disgust, anger, and neutral
- Using text as our input rather than audio (using TTS technology for converting text to audio)
- Adding head movement to the talking faces to make them more realistic
- Using evaluation metrics to compare to other methods and find areas of improvement

We are planning to continue the project that was done by other students last year, which means we will be using and improving the Wav2lip method for lip synchronisation. We are aware that there are issues with quality of lip movements and the absence of head movement. We are open to implementing other techniques to improve these aspects and produce

a more refined product. Moreover, we are aware that incorporating emotions into talking faces is challenging and there is relatively little research in this area. After executing last year's project, we have concluded that it has some imperfections and did not fully meet our expectations. Although we were able to differentiate between different emotions such as happy and sad, the resulting facial expressions were distorted and not representative of the original person. This year's objective is to resolve these issues and create realistic emotional talking faces that convey the intended emotions accurately with the head movement and conversion of text input to audio input.

3. Literature Review

As I have identified five key features, I have broken down the literature review into five parts, with the requirements listed above and the evaluation metrics.

3.1. 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. Numerous studies have been conducted on lip synchronisation, and various techniques have been proposed. Some of the research papers [1][2] propose methods using a generative adversarial network (GAN), which is composed of two neural networks, the generator which generates a single image using encoders and decoders, and the discriminator which distinguishes the generator's data from real data. GANs have the advantage of being able to generate new samples that are highly realistic, as they use random noise as input instead of relying on expected data distribution. However, this freedom also makes it difficult to control the output and ensure that it meets specific constraints or focuses on certain characteristics. [1] In [2], they propose a model called LipGAN, for generating realistic talking faces conditioned on audio in any language. The LipGAN algorithm employs adversarial training to improve the lip-sync accuracy of the videos it generates. Another model is proposed in [3], which is called Wav2Lip, a lip-synchronisation network, that is significantly more accurate than earlier studies for lip-syncing arbitrary talking face videos in the wild with arbitrary speech. Wav2Lip uses a standard encoder-decoder architecture that takes the target pose and target speech as input and generates a lip-synced face. [4] The paper [3] also proposes integrating GAN into the Wav2lip model. In contrast to the LipGAN architecture defined in [2], the discriminator used in this study is not further trained with the generator but is a pre-trained discriminator. [3] This method produces improved lip synchronisation

results. As we are aiming to improve last year's project which used the Wav2lip model, we have narrowed our focus on studies that either use Wav2lip or compare alternative methods or their own methods to it.

Some studies [4][5][6][7] identified the limitations of the Wav2lip model, comparing it with their own and other existing methods. In many frames generated by the Wav2lip model, the interior of the mouth appears blacked out, resulting in the loss of details such as the teeth. [5] Also, the Wav2Lip model generates videos at a relatively low resolution of 96 × 96 pixels [4] causing the mouth region to appear blurry and lack finer details. Furthermore, the Wav2lip model cannot render emotion [6] and is unable to obtain head poses from audio, resulting in fixed head positions in the generated videos. [7] These limitations can affect the realism of the talking face features in the generated videos, which need to be further improved for our project. In addition to the Wav2lip model, more recent studies (such as [7] and [8]) have been conducted on lip synchronisation and produced better results than the Wav2lip model. In [7], a one-shot talking face generation strategy is introduced which also uses encoder-decoder architecture and temporal discriminator using PatchGAN and Lip-sync discriminator which employs SyncNet in Wav2Lip. In [8], a model named PC-AVS is introduced which is a Pose-Controllable Audio-Visual System. Although PC-AVS is capable of performing lip synchronisation with a moving head, it has some limitations. These include its requirement for aligned faces and the resulting mechanical and unnatural appearance of the mouth shapes as it has difficulty producing consistent talking styles. [7] By expanding our scope to include these studies, we could gain a more comprehensive understanding of the most recent approaches and potentially incorporate their sights into our own project. Although the students who worked on last year's project were aware of the limitations of the Wav2lip model and attempted to resolve them by incorporating the Spatial Transformer model (SPT) and DeepFaceLab, the modifications did not fully meet our expectations. By comparing and adopting a range of approaches, we intend to produce the most effective strategy for lip synchronisation. In Table 1, the proposed models are listed with their advantages and disadvantages for easy comparison and evaluation with the Wav2lip model.

Table 1: Comparison of two recently proposed models to the Wav2lip for lip synchronisation

	Advantages	Disadvantages	Ref.
Wav2lip	• Accurate mouth and jaw movement	• Mouth region is blurry • Low resolution • No emotional expression • No head movement	[4] [5] [6] [7]
PC-AVS	• Moving head	• Unnatural lip movement • No emotional expression	[7] [8]
One-Shot	• Natural mouth lip movement • Moving head	• No emotional expression	[8]

3.2. Talking face generation with emotions

As previously mentioned, emotions are a vital component of human conversation, and we aim to incorporate emotions into the talking faces we generate. In [10], they propose a neural network system that generates emotional talking faces from speech conditioned on categorical emotions using the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) dataset. The CREMA-D dataset contains 7,442 video clips of 91 actors and actresses, (48 male and 43 female) of diverse age and ethnicity, expressing six categorical emotions which are anger, disgust, fear, happiness, neutral, and sadness. [9] The neural network proposed in [10] takes three inputs: a speech utterance, a reference face image, and a categorical emotion condition. Using this information, the network generates a talking face that is synchronised with the input speech and incorporates appropriate emotional expressions. [10] The emotion label is encoded as a one-hot vector and passed through an emotion encoder which utilises a two-layer fully connected neural network to convert the one-hot vector into an emotion embedding then in the emotion discriminator, probabilities of seven classes (six emotions listed above and whether if it is fake) is calculated out by using emotion GAN. [10] Although the model performs well on the trained dataset, it has limitations when it comes to generalising to unseen datasets as identified in [6]. The paper [6] and [11] propose an emotional talking face generation model called IJCAI and Emotion-Aware Motion Model (EAMM) respectively, using the Multi-view Emotional Audio-visual Dataset (MEAD) [12]. MEAD includes talking-face videos recorded by 60 actors and actresses talking with eight different emotions at three different intensity levels. [12] The emotion-guided optical flow-based texture deformation network in the model proposed in [6] generalises better for arbitrary target subjects by using one-shot learning. While [6] and [10] have only one module for the whole process, EAMM is composed of 2 modules, an Audio2Facial-Dynamics (A2FD) module that generates audio-driven talking face generation with neutral expressions and an Implicit Emotion Displacement Learner that involves emotional dynamics. [11] In order to incorporate emotional information into the A2FD module, an emotional extractor is used to extract the emotion feature from the processed video frames. [11] Then, to generate emotion dynamics that are synchronised with the input audio, the key-points and their corresponding jacobians predicted from the A2FD module, along with the emotion feature are used as inputs to the displacement predictor which utilises a 4-layer multiple layer perceptron (MLP) to predict the displacements for the key points. [11] The model produced an emotional accuracy of 58%, while the actual video had an accuracy of 71%. [11] Given that the EAMM [11] and IJCAI [6] appear to produce the most promising output for emotion incorporation, we aim to obtain further insights from them for our project. For incorporating emotions into talking faces, Table 2 lists the recently proposed models with their advantages and disadvantages.

Table 2: Comparison of three recently proposed models for adding emotions.

	Advantages	Disadvantages	Ref.
IJCAI	• Whole face encoder is used so the expression on the entire face is possible • Provide good generalisation for arbitrary target subjects	• No head movement, not completely realistic • Only one module for generating lip synchronisation and emotion so emotion cannot be modified independently	[6]
Model in [10]	• Whole face encoder is used so the expression on the entire face is possible	• No head movement • Dataset specific so does not generalise well in another dataset • Only one module for generating lip synchronisation and emotion so emotion cannot be modified independently	[6] [10]
EAMM	• Whole face encoder is used so the expression on the entire face is possible • Includes head pose (head movement) • Separate module for generating lip synchronisation and emotion so the emotion can be trained separately	• Still not completely realistic	[11]

3.3. Conversion of text to audio input

For the talking face generation, audio input can be used in which the system takes an audio clip as input and synchronises the movements of the lips with the spoken words. This method has been widely explored in numerous studies and resulted in creating realistic talking faces. In addition to audio input, text can be taken as an input as well and we will be using the text input in our project as we are planning to apply this to chatbot in the future. There were relatively few papers that took the text as an input rather than audio, but some papers [5][13][14] introduced TTS to convert text to audio and generate video by applying this technology. The TTS system converts an arbitrary ASCII text to speech. [13] The system analyses the text to identify its phonetic components, such as sound units, word boundaries, and prosody markers. Then it matches these phonetic symbols with the appropriate sounds stored in its inventory and combines them together to form the acoustic signal for the voice output device. [13] In simpler terms, the system converts the text into phonetic symbols and then uses those symbols to generate the audio output. Traditionally, the TTS system has employed a two-stage pipeline, the first stage of using an acoustic model to generate an intermediate speech recognition (i.e., mel-spectrogram) and the second stage of converting the speech representation to a raw waveform. [5] The studies showed the successful output of converting text to audio and applying it to generating talking faces. While TTS is a useful tool that reduces human workload, its emotionless nature [13] is a concern for our project. Despite our research, we were unable to find any relevant papers on incorporating emotions into the TTS system.

3.4. Head movement and resolution

The movement of the head is also an important factor for generating realistic talking faces since the face may appear unnatural without the head movement. Since Wav2lip lacks the ability to generate head movement on its own, we plan

to incorporate head movement techniques into our project. 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 the audio input. In a related paper [15], a deep-learning based architecture is proposed to predict facial landmarks, capturing overall head poses, from only speech signals while some papers [8][11] employ pose source videos to compensate only for head motions. PC-AVS which was mentioned in section 2.2., identifies learning pose motions from audio mostly keeps the original pose unchanged as there is a lack of absolute pose information that can be inferred from audio signals alone. [8] The head pose is encoded onto the target frame using data augmentation techniques. Furthermore, another paper [11] also notes the difficulty of the inferring head pose from audio signals alone and proposes incorporating the pose sequence estimated from a training video clip as an additional input to the model. A 6-dimensional vector (i.e., 3 for ration, 2 for translation, and 1 for scale) is used to represent the head pose for each frame in this paper. [11] Moreover, although Wav2lip is effective in generating accurate lip and jaw regions, the resulting videos lack detailed facial features such as teeth, lip colour, and face texture in the lower half of the generated face and it generates videos with a resolution of 96 x 96 pixels. [4] While Wav2lip achieves lip synchronisation by masking the lower half which results in leakage of mouth information encoded in the top half of the embedding, [4] proposes training a separate Pose-VQGAN model to avoid any unnecessary leakage of information. Teco-GAN is also used for high resolution and the model is trained with 4K Talking Face Dataset (4KTF) which includes 140 YouTube high-quality videos to produce high resolution output. [4] As part of our project, we have discovered some promising strategies through our research. Since we plan to incorporate the techniques for head movement and improving the resolution of the generated videos as our project, we will be trying to implement and improve these methods to achieve these objectives. For the head movement, Table 3 lists the recently proposed models with their advantages and disadvantages.

Table 3: Comparison of three recently proposed models for head movement

	Advantages	Disadvantages	Ref.
PC-AVS	• Can generate any head movement the user wants	• No emotional expression	[8]
EAMM	• Can generate any head movement the user wants	• Not found yet for head movement and also provides emotional expression	[11]
MakeItTalk	• If generated well, the head movement can be extracted from the audio	• Sometimes experience difficulty of inferring head pose from audio signals alone • No emotional expression	[11] [15]

3.5. Evaluation metrics

Evaluation is essential for talking face generation in order to evaluate the quality and efficacy of generated videos and for comparing different models and identifying the optimal strategy for a given task. For evaluating the generated videos, we decided to measure the image quality, accuracy of lip synchronisation, and accuracy of emotions for our project. Peak SNR (PSNR) and Structural Similarity (SSIM) are used between the generated video frames and the ground-truth video frames to evaluate the image quality of the generated videos. [6][10] In [6], CPBD and FID metrics are used additionally to measure image quality. Landmark Distance (LD) and Landmark Velocity Difference (LVD) are used to evaluate the landmark quality. To quantify the accuracy of lip displacements M-LD and M-LVD are used and F-LD and F-LVD are used to measure the accuracy of facial expressions and head pose. [6][11][16] For the audio-visual synchronisation which also means lip synchronisation is measured using the normalised landmarks distance (NLMD) between landmarks extracted from the generated and ground-truth video frames. [10] The emotion classifier network in EVP [16] is used in [6] to measure emotion accuracy and SyncNet to estimate audio-visual synchronisation. Qualitative evaluation of the generated videos and other state-of-the-art methods is also presented in [6][11]. They trained their model and other proposed techniques such as MakeItTalk [15], EVP [16], and Wav2lip on the same dataset and made a comparison. They also made a comparison with the ground truth videos (real videos), which allows researchers to identify areas where the system's performance can be enhanced and optimised. For our project, we are planning to conduct both quantitative and qualitative evaluations. By conducting both types of evaluation, we aim to acquire a more complete understanding of our proposed method's strengths and weaknesses and to identify areas for improvement.

4. Project Scope

The goal of our project is to modify and improve previously developed technologies for the generation of talking faces with emotions. After completing the literature survey, we realised that there is currently no perfect solution for generating natural lip synchronisation and for incorporating emotions into talking face generation. While some studies have made significant progress in these areas, there is still a lot of work to be done for generating realistic and accurate models. Furthermore, the literature survey has revealed that we require a technique for converting text input to audio input. We also plan to integrate and develop the methods proposed in our literature review and use both quantitative and qualitative evaluation methods to compare our proposed method to other existing methods and to find areas of further improvement. We will be using our knowledge gained from the literature survey to generate a natural lip synchronisation and emotional expression with head movement and text to audio conversion. Our planned implementation is shown in Figure 1 below, which illustrates the brief pipeline we will use. To achieve our goals effectively, we have divided our roles and

responsibilities for a specific period of time. The Gantt chart is shown in Figure 2 below. We began by conducting a literature review and analysing last year's project to identify areas for improvement. We now intend to start developing and testing our emotional talking face generation. Our goal is to complete this stage of the project by the end of semester break, to allow us more time to focus on the display and writing the final report.

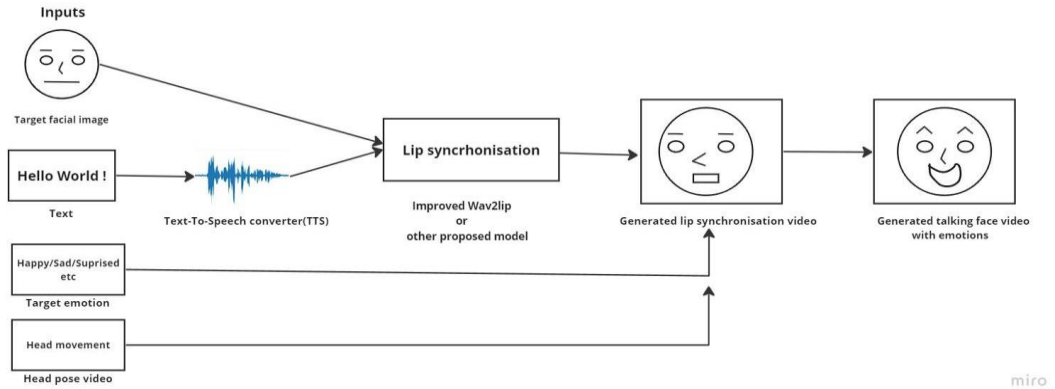

Fig. 1. Our general pipeline on our development

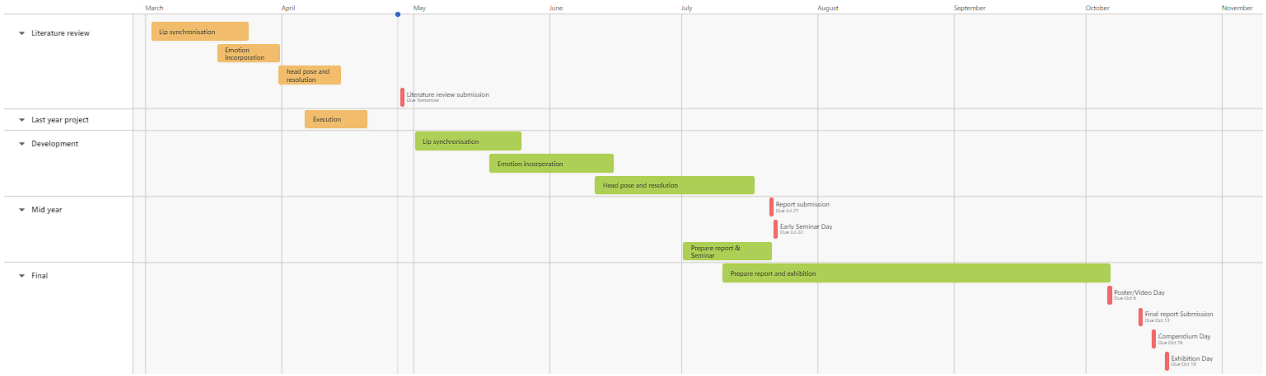

Fig. 2. Gannt chart

5. Conclusions

In conclusion, through our literature review, we have identified a number of key features for our project, which are lip synchronisation, talking face generation with emotions, conversion of text to input audio, and head movement and resolution. This review has provided us with valuable insights into the existing methods and helped us understand the areas that need to be developed throughout the project. Even though several studies made significant progress in developing talking face with emotions, there is still much space for development. Moreover, our literature review has given us important knowledge regarding the various evaluation metrics that can be used to compare and improve our model. By using knowledge gained from the literature review, we aim to develop a system that generates a realistic and accurate talking face with emotions.

Acknowledgements

I would like to thank the supervisor Ho Seok Ahn and Trevor Gee for providing their previous project and guidelines and support that will be further needed for this project.

References

[1] 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

[2] Prajwal K R, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, C. V. Jawahar, "Towards Automatic Face-to-Face Translation", Mar. 2020, Available: https://arxiv.org/abs/2003.00418

[3] 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, Available: https://arxiv.org/abs/2008.10010

[4] Anchit Gupta, Rudrabha Mukhopadhyay, Sindhu Balachandra, Faizan Farooq Khan, Vinay P. Namboodiri, C. V. Jawahar, "Towards Generating Ultra-High Resolution Talking-Face Videos with Lip synchronization", 2023, IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

[5] Hyoung-Kyu Song, Sang Hoon Woo, Junhyeok Lee, Seungmin Yang, Hyunjae Cho, Youseong Lee, Dongho Choi, Kang-wook Kim, "Talking Face Generation with multilingual TTS", 2022 IEEE/CVF Conference on Computer Vision and Pattern recognition (CVPR)

[6] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, Brojeshwar Bhowmick, "Emotion-Controllable Generalized Talking Face Generation", May. 2022, Available: https://arxiv.org/abs/2205.01155

[7] Suzhen Wang, Lincheng Li, Yu Ding, Xin Yu, "One-shot Talking Face Generation from Single-speaker Audio-Visual Correlation Learning", Dec. 2021, Available: https://arxiv.org/abs/2112.02749

[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, Available: https://arxiv.org/abs/2104.11116

[9] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset", 2014, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 5, NO. 4

[10] Sefik Emre Eskimez, You Zhang , Zhiyao Duan , "Speech Driven Talking Face Generation From a Single Image and an Emotion Condition", 2022, IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 24

[11] 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", 2022, Available: https://arxiv.org/abs/2205.15278

[12] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, Chen Change Loy, "MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation", Available: https://wywu.github.io/projects/MEAD/MEAD.html

[13] Dr. S.A. Ubale, Girish Bhosale, Ganesh Nehe, Avinash Hubale, Avdhoot Walunjkar, "A Review on Text-to-Speech Converter,"June 2022, IJIRT volume9 issue 1.

[14] Sibo Zhang, Jiahong Yuan, Miao Liao, Liangjun Zhang, "TEXT2VIDEO: TEXT-DRIVEN TALKING-HEAD VIDEO SYNTHESIS WITH PERSONALIZED PHONEME - POSE DICTIONARY", 2022, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[15] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, Dingzeyu Li, "Make it talk: Speaker-Aware Talking head animation", 2021, Available: https://arxiv.org/abs/2004.12992

[16] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, Feng Xu, "Audio-Driven Emotional Video Portraits" , IEEE/CVF Conference on Computer Vision and Pattern Recognition