DeepEmotion: Emotional Audio-Driven Talking-Cartoon

Ionut Bostan 6645704
Bachelor of Science in Computer Science
Dr Suparna De

08 November 2022

1 Introduction

Artificial intelligence (AI) is changing how we interact with our daily activities through the development of smart gadgets and numerous intelligent systems. Such a device has to be able to help people by determining their needs and reacting to their emotions.[1] Due to their ability to immediately transmit the speaker's intention through facial movements, emotional expressions can be extremely important in human-computer interaction. As a result, I want to provide an all-encompassing method that creates a lifelike talking-cartoon animation that conveys spoken emotions, i.e., amusement, confusion, anger, and excitement. An actual face photograph and a passage of text will be used as the basis for the audio-driven talking cartoon. The following are some examples of use cases for a talking avatar: Voice assistant chatbots like Amazon Alexa, Google Assistant, or Siri, virtual worlds, teaching, or, in the case of current research, "Risk Playgrounds," which allow users to safely experiment with potentially risky interactions [2]. Photorealistic cartoon avatars, according to research, improve the student learning experience in an online setting and provide a more stimulating environment for teacher-student interaction in a higher education context.[3]

1.1 Aims

- Create an end-to-end system that generates a talking cartoon utilising state-of-the-art image-to-image translation and audio-driven character models. For the voice synthesis, I propose a text-to-speech (TTS) transformation approach that leverages sentiment-based text analysis to make the avatar sound more lifelike.
- As a stretch goal, create and launch a web application based on the generated model as a "proof of concept".

1.2 Objectives

- Acquire emotional text and voice data. Preprocess the text data and create a model that successfully classifies sentiments. Investigate, evaluate, and create a Text-To-Speech model that employs sentiment-based text analysis as a front end. Then, retrain the acoustic component of this model using emotional voice data to generate an emotionally synthesised voice. Finally, check all the inputs/outputs of all components in the talking-cartoon pipeline, and make any necessary changes to establish an end-to-end system.
- Examine all of the work completed thus far at the start of the second semester and determine whether the second aim of deploying the model is doable in the remaining time.
- Knowledge gathered in last year's modules for Information Retrieval (COM2034) and Artificial Intelligence (COM2028), as well as this year's modules for Computational Intelligence (COM3013) and next semester's Natural Language Processing (COM3029), will aid me in accomplishing my objectives.
- Gain knowledge on new topics such as Aspect-Based Sentiment Analysis (ABSA) and necessary Deep Learning technologies. Develop new skills by learning necessary frameworks.
- Make an accurate work schedule and account for unforeseen circumstances that may cause my workflow to be disrupted. Follow the timetable as closely as possible.

2 Literature Review

For a long time, cartoon avatars have been employed in numerous areas such as education, computer games, entertainment, and video streaming; yet, making these types of animations remains a significant issue[5][7], particularly when a cartoon is used instead of a genuine human face image. This section will introduce each component of the proposed system. The Text-To-Speech (TTS) Model is divided into three parts: the text analysis front-end, the acoustic module, and the voice synthesis module. Finally, there is an Image to Image Translation model and an Image Animation model.

2.1 Sentiment Analysis in Text

Sentiment analysis is the use of Natural Language Processing (NLP) to identify emotions from the text. The text analysis aims to categorise emotions such as angry, sad, happy or positive/negative polarities using a spoken word and the context around it[4]. The emotion classification method can be split into two main categories:

- Classic classification algorithms such as Bag of Words (BoW)[6], Support Vector Machine (SVM), Decision Tree, Naive Bayes or a linear classifier such as Logistic Regression.
- Deep Learning approaches using Deep Neural Networks such as Long Short Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Convolutional Neural Networks CNN and more recently, Generative Adversarial Networks (GAN).

2.1.1 Acoustic module

The preceding block's linguistic features, such as phonemes, phonemes duration, and pitch, are combined with the speaker embeddings in this section to create an acoustic feature/Mel-spectrogram.

2.1.2 Voice Synthesis module

The acoustic feature/Mel-spectrogram is converted into a waveform using this module. A Neural Network may also be used to learn the mapping from the Mel-spectrogram to the waveform.

2.2 Image to image translation

The goal of image-to-image translation is to learn how to map an input image to an output image using a loss function in order to change an image from one domain to another. The concept of Generative Adversarial Networks[10] (GANs) was initially presented by (Goodfellow et al., 2014)[10] and is one of the most used architectures for domain-adapted image generation. To acquire the domain-adapted cartoon picture in my system, I'll be employing a cutting-edge StyleGAN[11] inversion framework.

2.3 Voice conversion

Because the synthesised version must keep the emotions from the sample speech, this part is critical to my proposed approach. The existing MakeItTalk[5] model, which will serve as the foundation of my design, uses a simple zero-shot speech conversion network. The network is made up of two encoders, one for producing content embedding from speech and the other for producing speaker embedding. Both encoders are then sent into a decoder, which generates the synthesised speech.[12]

2.4 Image Animation

This is the system's last stage, in which a video of an audio-driven talking cartoon created. The model can generate both face expressions and head movements from the supplied audio.[5] Facial landmarks are inferred from the input image utilising Estimation-Correction-Tuning (ECT)[13] in training by applying creative image augmentation[8]. The disentangled audio material then drives head movements and face landmarks.[5]

3 Technical overview

The pipeline in Figure 1 seeks to generate a talking cartoon that can express the emotion from the text into its voice given a textual input and an image. My primary contribution will be to the Text-To-Speech (TTS) module, namely, in the sentiment analysis of the provided text. The initial objective will be to find a well-balanced emotional text dataset, as most review datasets have a stronger positive polarity. For the Aspect-Based Sentiment Analysis (ABSA), a Deep Neural Network (DNN) model could be employed to analyse the input text to determine the emotional tone of the given text input. An alternative method would be a classical NLP method that will require more data processing, such as tokenisation, stopword removal, and steaming. Following the transformation of text into a waveform, the next step is to create the avatar from the input image. For this task, I will use a state-of-the-art HyperStyle.[11] which uses a Generative Adversarial Network (GAN) Inversion to generate a domain-adapted image. The newly generated image will have to be resized to 256x256 pixels and, together with the synthesised audio, is fed into a speaker-aware talking-head animation. For this task, I have chosen the MakeItTalk[5] model. Facial landmarks are extracted from the cartoon, and the input audio is passed into a voice conversion model that extracts the content and speaker identity embeddings. Finally, a video of the cartoon speaking the input text will be generated.

Some tests were conducted in order to construct a talking cartoon using a piece of text and an image of my portrait. This was performed using the following methods: I used a Google Colab notebook to run the pre-trained HyperStyle[11] model to generate the cartoon picture. Then I utilised a web API[9] for text-to-speech conversion. Because the picture-to-cartoon model outputs a 1028x1028 image size, the resultant cartoon image needs to be scaled to 256x256 pixels. The final talking cartoon was generated using the MakeItTalk[5] model in another Colab notebook, having the newly generated cartoon image and sound as input.

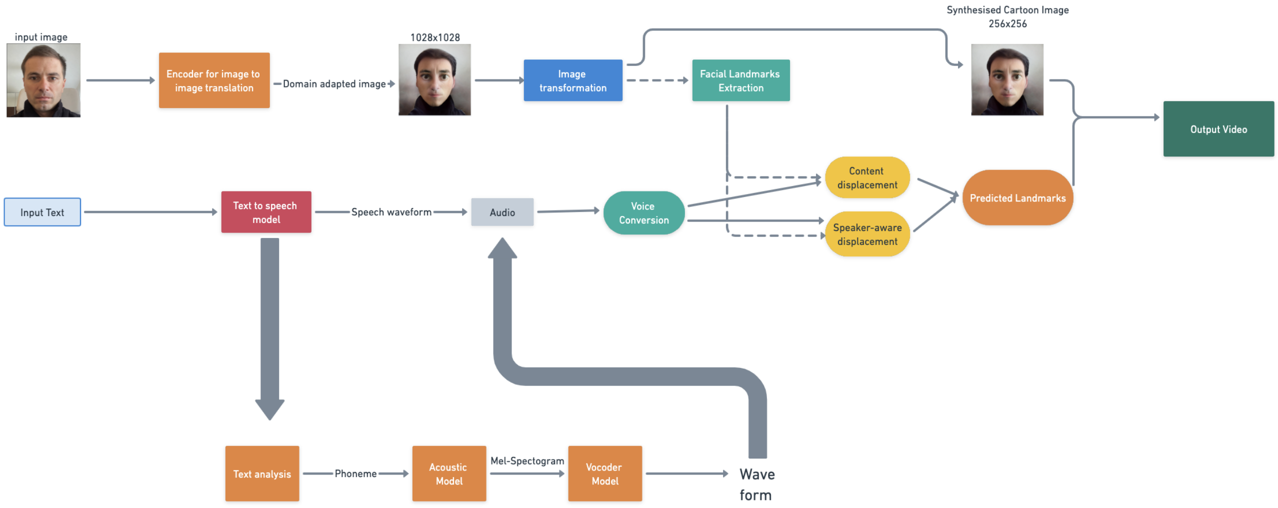


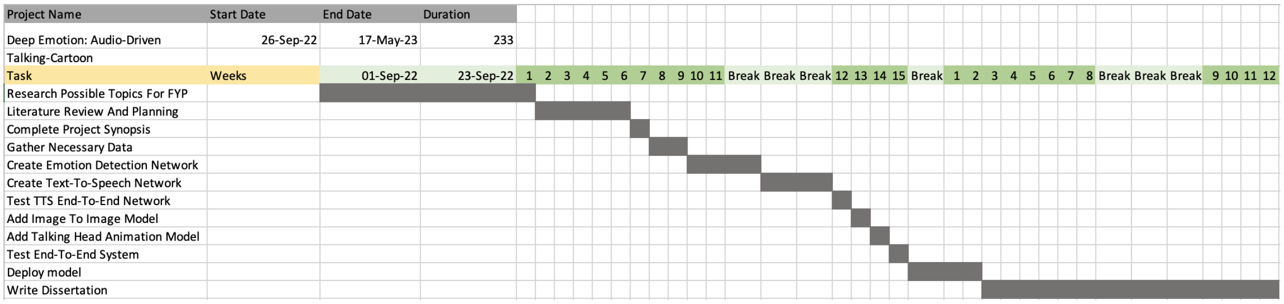
Figure 1: Talking-Cartoon Pipeline

4 Workplan

The following work plan is what I will be using for the project is shown in Figure 2

4.1 Risks

- Unable to complete the work due to unforeseen circumstances
- The project scope is too ambitious and necessitates skills that are not obtainable in the period available
- Integration of existing models or APIs is not feasible
- Poor planning and management of time
- Licensing issues.



Project Name	Start Date	End Date	Duration																																		
Deep Emotion: Audio-Driven	26-Sep-22	17-May-23	233																																		
Talking-Cartoon																																					
Task	Weeks	01-Sep-22	23-Sep-22	1	2	3	4	5	6	7	8	9	10	11	Break	Break	Break	12	13	14	15	Break	1	2	3	4	5	6	7	8	Break	Break	Break	9	10	11	12
Research Possible Topics For FYP		■	■	■																																	
Literature Review And Planning					■	■	■	■	■																												
Complete Project Synopsis										■																											
Gather Necessary Data											■	■																									
Create Emotion Detection Network													■	■	■																						
Create Text-To-Speech Network																■	■																				
Test TTS End-To-End Network																		■																			
Add Image To Image Model																			■																		
Add Talking Head Animation Model																				■																	
Test End-To-End System																					■																
Deploy model																						■	■	■													
Write Dissertation																									■	■	■	■	■	■	■	■	■	■	■	■	■

Figure 2: Project Timeline

References

[1] Chairs Constantine Stephanidis, Gavriel Salvendy, Members of the Group Margherita Antona, Jessie Y. C. Chen, Jianming Dong, Vincent G. Duffy, Xiaowen Fang, Cali Fidopiastis, Gino Fragomeni, Limin Paul Fu, Yinni Guo, Don Harris, Andri Ioannou, Kyeong-ah (Kate) Jeong, Shin'ichi Konomi, Heidi Krömker, Masaaki Kurosu, James R. Lewis, Aaron Marcus, Gabriele Meiselwitz, Abbas Moallem, Hirohiko Mori, Fiona Fui-Hoon Nah, Stavroula Ntoa, Pei-Luen Patrick Rau, Dylan Schmorrow, Keng Siau, Norbert Streitz, Wentao Wang, Sakae Yamamoto, Panayiotis Zaphiris & Jia Zhou (2019) Seven HCI Grand Challenges, International Journal of Human–Computer Interaction, 35:14, 1229-1269, DOI: 10.1080/10447318.2019.1619259.

[2] AP4L: Adaptive PETs to Protect and emPower People during Life Transitions. https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/W032473/1 (accessed Sept. 30, 2022)

[3] P. Craig, N. Roa Seiler, A. Benitez Saucedo, M. Martinez Diaz, J. Castañeda Santos, F. Lara Rosano (2014) THE ROLE OF PHOTO-REALISTIC AND CARTOON AVATARS IN A BLENDED-LEARNING ENVIRONMENT, EDULEARN14 Proceedings, pp. 303-308.

[4] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi and Puneet Agrawal 2019. SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. Microsoft, India anchatte, kedharn, mejoshi, punagr@microsoft.com

[5] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. MakeItTalk: Speaker-Aware Talking-Head Animation. ACM Trans. Graph. 39, 6, Article 221 (December 2020), 15 pages. https://doi.org/10.1145/3414685.3417774.

[6] Doaa Mohey El-Din. Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis 2016. Information Systems Department Faculty of Computers and Information, CU Cairo, Egypt.

[7] Deepali Aneja, Wilmot Li. Oct 2019. Real-Time Lip Sync for Live 2D Animation. https://github.com/deepalianeja/CharacterLipSync.

[8] Jordan Yaniv, Yael Newman, Ariel Shamir. ACM Transactions on Graphics, accepted, 2019. The Face of Art: Landmark Detection and Geometric Style in Portraits.

[9] Text to speech online. https://speechify.com (accessed Oct. 12, 2022).

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair†, Aaron Courville, Yoshua Bengio. Generative Adversarial Nets June 2014. Departement d'informatique et de recherche op ´ erationnelle ´ Universite de Montr ´ eal ´ Montreal, QC H3C 3J7.

[11] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, Amit Bermano. HyperStyle: StyleGAN Inversion with HyperNetworks for Real Image Editing 2022. Blavatnik School of Computer Science, Tel Aviv University.

[12] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Mark Hasegawa-Johnson. AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. Jun 2019.

[13] H. Zhang, Q. Li, Z. Sun, and Y. Liu. 2018. Combining Data-driven and Model-driven Methods for Robust Facial Landmark Detection. IEEE Transactions on Information Forensics and Security 13, 10 (2018), 2409–2422.