Enhancing AI Powered NPC Facial Animations by Training SadTalker*

Extended Abstract†

Harshal Mishra
a1888203@adeliade.edu.au

Supervisor: Dr Zhibin Liao

ABSTRACT

The Progress of artificial intelligence in non Player charters has enhanced user experience in Games. Realistic facial animations are crucial for immersive interactions, conveying emotions. This Project explores the potential of SadTalker which is a face generation model to improve AI powered NPCS by creating complex and emotionally expressive animations. This project conducted an experiment by training SadTalker on a dataset and compared the results with the pre trained model. Then through the use of psychological models of emotional recognition a qualitative analysis of the generated animations to assess their ability to display complex emotions. The results showed the limitations of training with scarce data and the showed the importance of having data diversity in training facial animations models.

1 INTRODUCTION AND MOTIVATION

In the realm of video games non player characters play a vital role in the user's experience. The realism and emotional depth an NPC can offer can impact player emersion and engagement in the game. Traditionally NPCs often suffer from very limited facial expressions and repetitive animations which often breaks the immersion of a living breathing world. Advancements in the field of Artificial intelligence have opened new inroads to solving this problem. This is done by leveraging models that are capable of generating realistic facial animations driven by audio inputs. This allows for the creation of NPCs that have dynamic expressions that reflect a wide range of emotions. SadTalker is one of such model.

The primary objective of this project is to explore how SadTalker can be improved to generate more complex and emotionally rich facial animation for AI powered NPCs. SadTalkers performance will be assed between the pretrained model and the one trained by a chosen dataset. The generated facial animations will be judged using psychological models of emotional recognition determine the complexity of emotions displayed. By identifying limitations and potential improvements the project aims to enhance the capabilities of SadTalkers Application in NPC development. This project contributes to the field by providing a detailed analysis of SadTalker performance between the pretrained dataset compared to the one trained using the CelpQ dataset.

2 PROBLEM DEFINTION

2.1 Difficulty in Understanding and Replicating Human Emotions

AI programs currently struggle with understanding human emotions, which is essential for generating authentic emotional expressions. The aspects that make human emotions so complicated are that they are multilayered and lay at the cross section of context, culture, and individual experiences. AI programs are likely trained on simplified models of human emotions such as the basic emotion theory, which was proposed by Ekman. This model only categorizes emotions into six buckets: happiness, sadness, fear, anger, surprise, and disgust (Ekman, 1992).

These types of models are incredibly useful for identifying broad emotional states; however, they fall well short of mapping the subtlety and variability of real human emotions. For example, the emotion "anger" can have vastly different manifestations depending on the context, such as ranging from silence to erupting in fury. These AI models would fail to distinguish between these two emotional states (Krumhuber et al., 2013). Additionally, AI programs struggle with the temporal dimension of an emotion, such as how emotions can change over time in response to different events. An example of this is how the AI in this paper failed to change emotional states after both attempts and escalations and de-escalation (Krumhuber et al., 2013).

2.2 Requirement of Large and Diverse Data Sets

The creation of AI NPCs that can respond to human emotions is heavily dependent on large datasets that encompass a wide range of human expressions for the AI to be trained on (Cowie et al., 2011). However, acquiring these datasets presents unique challenges.

The collection of emotional datasets is incredibly resource intensive. These datasets require annotating a vast array of emotional expressions across different contexts, cultures, and individual differences. They also require not only facial expressions but also vocal tones, body language, and contextual information that help interpret how emotions are understood. As noted by Paier et al. (2021), the creation of just one single actor video corpus for training neural networks is a laborious process that requires planning to ensure data is usable for training purposes.

This process of annotation creates its own set of problems, as annotations are required to understand the nuanced emotions being expressed in the moment. However, due to human biases and inconsistencies, especially in complex emotional situations where distinctions between certain emotions are difficult to ascertain, the accuracy and reliability of these datasets become diminished.

2.3 Difficulty in Creating Nuanced Emotional Expressions

Generating nuanced emotions presents a monumental challenge for AI. This is due to the fact that emotions are not binary and land on a spectrum where humans express their relative position on said

spectrum through very subtle cues. The ability to produce these subtle and nuanced expressions is critical to creating an NPC that can be believable enough to resonate with the player, allowing for deeper immersion (Krumhuber et al., 2013).

A technical challenge that presents itself is the AI's ability to make subtle variations in facial muscle movements that accurately correspond to different emotional states (Paier et al., 2021). This is because the current standard for facial animations set by the Facial Action Coding System (FACS) is limited in its ability to map complex, nuanced facial expressions. FACS does provide a framework for categorizing facial expressions; however, it doesn't provide enough detail to capture the interplay between small facial muscles that impact the context of an emotional response (Ekman & Friesen, 1978**).**

3 LITERATURE REVIEW

3.1 Example-Based Facial Animation of Virtual Reality Avatars Using Auto-Regressive Neural Networks

This paper written by Paier dives into how a hybrid approach to facial animation that combines both auto-regressive neural networks with example-based methods to create realistic facial animations for virtual reality avatars. The methodology tackles the third problem outlined as it creates nuanced emotion expressions by leveraging an auto-regressive neural network.

3.2 Synthesizing Obama: Learning Lip Sync from Audio

Suwajankorn explored a method of synthesizing lip sync from audio input with a focus on creating realistic facial animations for former U.S. President Barack Obama. This was done by training a neural network to predict mouth shapes based on audio data. This paper explores the second problem as even with a large target dataset for an individual, there were several problems with the AI limiting its expressions.

3.3 Effects of Dynamic Aspects of Facial Expressions: A Review

This review focuses on the importance of dynamic aspects of perception in recognition of facial expressions. This paper highlights how dynamic facial expressions, in contrast to static images, drastically increase the accuracy of emotional recognition and increase the perceived intensity and authenticity of the emotions on display. This relates to the first problem and how understanding human emotions has a temporal dimension.

3.4 Facial Emotion Expression Corpora for Training Game Character Neural Network Models

This paper explores the creation and validation of facial emotion expressions for the sole purpose of training neural networks to drive NPC facial animations. The authors use real actors to help the model generate life-like animations. This paper addresses both the first and second problem. It explores the requirements for large datasets and the difficulty in replicating real nuanced human emotions. The authors also discuss a potential solution, suggesting training NPCs on specific datasets, reducing the size of the dataset while increasing the accuracy of emotion replication.

3.5 Measuring Emotion Velocity for Resemblance in Neural Network Facial Animation Controllers and Their Emotion Corpora

This paper explores a unique approach for evaluating the accuracy of a neural network. It suggests that by measuring the velocity of emotional expressions and comparing them to human actor performances, the temporal dynamics of expressions are preserved in animations. This methodology is critical as it provides a potential solution for the third problem by creating nuanced emotional expressions that change over time.

4 METHODOLOGY AND EXPERIMENTAL SETUP

4.1 Data Preparation

The images were collected from the Celea dataset which is a collection of over 200,000 high resolution facial images depicting a series of emotions. These images were pre-processed to meet SadTalkers input requirements which included resizing all images to a consistent dimensions and normalisation of pixel values between -1 to 1. The images were stored in a directory within the SadTalker project. A single audio file containing a recorded monologue of a Video game character which went through several emotions was selected to drive the facial animations and was also placed in the project directory.

To capture the essential facial features in a image necessary for animations a 3d morphable model coefficient was created using the SadTalkers preprocessing scripts. These coeffects are crucial for the model to understand the image and identify the faces and its landmarks. Then train.txt and val.txt files were created each containing the identifier of the single audio sample and images and placed int eh projects file list directory.

4.2 Configuration adjustments

The configuration files were adjusted to meet the custom dataset. In the audio2exp.yaml file that controls the auto to expression component of SadTalker the variables were updated so the data paths point to the custom images and audio. The batch size was set to one due to minimal data availability in the audio department. Similar adjustments were made to the audio2pose_unet_noAudio.yaml file for the audio to pose model to ensure that all variables that referenced to data sources matched with the correcting pathing.

4.3 Training Process

Then all configuration files were ran to train SadTalker with the new custom dataset.

5 RESULTS

Pretrained model animation can be viewed here:
https://youtube.com/shorts/jtUxLS6pqZ8?feature=share

Custom Dataset trained model animation can be viewed here:
https://youtube.com/shorts/K-wCcY4580I?feature=share

5.1 Qualitative Assessment Using Psychological model

To evaluate the complexity and authenticity of the facial animations the facial action coding system [Ekman and Friesen, 1978] and Ekman's basic emotions framework [Ekman, 1992] which are widely recognized psychological model for emotional recognition and analysis.

Custom trained model

The animations generated by this model showed minimal changes in facial expressions primarily remaining more neutral regardless of the emotional context in the audio. Acorrding to the FACS model the actions unis corresponding to the specific muscle movements were not demonstrated. A major recuring example was the lip corner puller during segments of the audio that convey happiness. Eye and eyebrow movements such as AU1 which references the infer brow raiser and AU5 which references the upper lid raiser which are crucial muscles for conveying expressive surprise and or fear (more surprise in the context of the audio clip) were largely absent. In totality the model failed to display complex emotions often defaulting to a neutral or ambiguous expression leading to large moments of emotion incongruence between the audio and the facial animations.

Pre Trained SadTalker model

In sharp contract the pre trained SadTalker models animations showed a wider range of facial expressions that corresponded to the audio emotional cues. The model effectively activated relevant Aus for different emotions. For example, the AU12 was activated during happy segments and AU15 that corresponds to the lip corner depressor for sad tones. More subtle eyebrow movement throughout the clip enhancing the realism of the model. Facial expression also changed more fluidly through the different emotional beats of the audio providing a more coherent experience.

6.2 Comparative Analysis

This time using Ekman's basic emotions framework the different emotions were assessed. The custom trained model primarily displayed neutral expressions. Meanwhile the pre trained model was able to convey four Surprise, Sadness, happiness and anger. The pre trained model also included more natural head movements such tilts and nods enhancing the realism where the custom trained model head remained much more static. Lip synchronization was also very poor in the custom trained model which violates the principle of phoneme-viseme correspondence which impacts realism of the model.

6 DISCUSSIONS

6.1 The impact on AI powered NPCs

The ability of NPCs to express complex facial emotions enhances user engagements and the pre trained SadTalker model demonstrated superior performance in generally all related subfields of this task. The custom trained model's shortcoming are highlight in the results due to the challenges of training with minimal data emphasising the importance of extensive and diverse datasets. So in conclusion the custom train dataset SadTalker model should not be used to power NPC facial animations for video games as It would lower engagement and emersion.

6.2 Implications and Contributions

This study contributes to filed by providing a analysis of SadTalkers performance when trained with a custom large dataset versus the pre trained model. It highlights the mission critical role of data diversity and volume in facial animations models capable of creating expressive and complex emotions. The use of psychological models offered a structured approach to facial animations evaluation and could be adopted in future studies and research. The finding suggest that developers should stick to using pretrained models rather than using more custom datasets which may not have sufficient data leading to unsatisfactory results.

7 LIMITATIONS

This project may provide some key insights for this field however it has major limitations. First the dataset having only one source of audio meant it lacked majorly in its data pool. The results would likely change with larger data sets. Second even though psychological models were applied for analysis these models are based on human interactions and faces meaning they do no accurate map on to NPC facial animations. So in the future a more qualitative source of evaluation may be more appropriate.

8 CONCLUSION

This project demonstrates that while custom training of SadTalker with minimal data is not practical for generating complex emotionally rich datasets the pre trained model fair pretty well in the psychological model evaluations. The methodology employed used is sound and easily replicable and invite clear areas for adjustments to configurations for the data to potentially induce a better result and further testing to be done.

9 FUTURE WORK

To advance the applications of SadTalker for AI powered NPC for future works should focus on the fine tuning of pre trained models with perhaps more domain specific data such as images of specific games to enhance personalisation, Implementing a attribute to the images of emotions and being able to generate specific animation of certain emotions independent of the audio clip would be despicable tool for future testing.

10 LINK TO GITHUB

This links to the github repository that holds the project
https://github.com/Harshalhellow/Sadtalker-project.git

The dataset for Celba can be found here
https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

REFERENCES

[1] Araujo, P. M., Reis, L. P., & Lau, N. (2020). Affective computing in digital games: Emotions and influence on game experience. *Journal of Multimodal User Interfaces*, 14(3), 271-281.

[2] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schroder, M. (2011). 'FEELTRACE': An instrument for recording perceived emotion in real-time. *Proceedings of the 9th European Conference on Speech Communication and Technology*, 1655-1658.

[3] Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4), 169-200.

[4] Ekman, P., & Friesen, W. V. (1978). Facial Action Coding System: A technique for the measurement of facial movement. *Consulting Psychologists Press*.

[5] Kappas, A. (2010). Smile when you read this, whether you like it or not: Conceptual challenges to affective computing. *IEEE Transactions on Affective Computing*, 1(1), 38-41.

[6] Krumhuber, E., & Scherer, K. R. (2011). Affect bursts: Dynamic patterns of facial expression. *Emotion Review*, 3(4), 442-443.

[7] Krumhuber, E., Kappas, A., & Manstead, A. S. R. (2013). Effects of dynamic aspects of facial expressions: A review. *Emotion Review*, 5(1), 41-46.

[8] Paier, W., Hilsmann, A., & Eisert, P. (2021). Example-based facial animation of virtual reality avatars using auto-regressive neural networks. *IEEE Computer Graphics and Applications*, 41(4), 47-58.