Affect Gesture Generation and Embodied Conversational Agents

Abdullah Ghanima
CS 549 - Artificial Intelligence for Visual Computing
Rutgers University, Department of Computer Science
Spring 2022

I. Abstract

An accurate simulation of a personality driven avatar has the potential to be extremely beneficial to the scientific community. On the project Embodied Conversational Agents, I have been working with Jerry Chang on creating an AI virtual training tool tasked with improving the skills of our teachers, with the intention of improving economic mobility in underprivileged neighborhoods. I first worked on implementing a basic conversational tool, which took speech as input and computed audio output through the use of natural language understanding (NLU).

I then shifted focus to the graphical components of the project, working with Artificial Intelligence with regard to visual computation. My work is related to the upper body gestures of a virtual agent. I first implemented an existing model, and from there I have been editing it in order to derive emotion and custom text to speech. I also carried out an in-depth statistical study to determine the accuracy of my assertions of emotion in a user's upper body gestures. This all has the purpose of creating a real life conversational agent.

II. Introduction

Arguably, one of the best ways to improve economic mobility in underprivileged neighborhoods is to improve the quality of the teachers. We will be creating a virtual training tool that has the purpose to better equip these teachers to handle difficult students and react appropriately. Our approach is to create an interactive artificial intelligence that will place teachers in some hypothetical scenario. The scenario will alter and continue based on the teacher's choices, and will conclude with a formal evaluation.. We have been working to create an AI embodied conversational agent, which will have emotional speech and upper body gestures. Creating such a program can have numerous other applications, since human behavior and action can be simulated. This is the description at a high level, and is the future intention of our work. My work focuses on the affect upper body gestures of a user. I have accurately determined seed sequences that result in the representation of a given emotion, allowing our AI to display their emotion via their gestures, I have also implemented an affect Text to Speech, which is another mechanism to display emotion.

III. Related Work

A. *Speech2AffectiveGestures [1]*

This is the main work that I have been researching with. I began by understanding their program, what the required inputs and outputs are. It takes in speech, text, seed sequence, and a seed pose. From these, it outputs a video of some set of upper body gestures displayed as vectors (10 vectors, each connected and sitting in a 3d space) for some number of frames depending on the video length. The text is displayed in the background, and the audio is playing. This program uses several well known technologies in the field of AI, and uses these to synthesize some set of upper body gestures given the text and audio. The source of its data is two datasets, the TED Gesture Dataset and the Trinity Dataset. They both contain videos of presenters, where they are creatively active in their upper body gestures. From this, they conducted extensive work in audio analysis, to extract emotion and tone through the use of MFCCs. They also used NLP to interpret the language itself. They also realized that each speaker had a different "personality" in their work. However, the program they created does not bound any of these variables together, through reworking with their code it would become possible to vary the input to better understand and represent these upper body gestures. Although this program is able to synthesize the upper body gestures, it currently does not accept custom speech or audio. It is also not conversational. In addition, it is not affect. My work focuses on these aspects.

B. *A Conversational Agent Framework with Multi-modal Personality Expression [3]*

There has however, been work that incorporates emotion. This work focuses on the emotion recognition and interpretation of an agent. However, it is not conversational and does include the language generation. Previously I worked on implementing NLU RASA, which is a natural language understanding tool, commonly used for conversational AIs. With my work, I intend to also incorporate the conversational portion into the upper body gestures of the embodied agent.

C. *M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues [2]*

Aniket Bera is a researcher who has extensive work with regards to emotion identification. He has worked on text as well as audio. Identifying the emotion of an agent is very difficult, but through attempts at quantifying human behaviour he has made significant progress in this field. I hope to work to add on to this, by adding upper body gestures and their corresponding emotions to the agents. This will not only add humanity to them, but also emphasize a given emotion if desired.

IV. Technical Approach and Results

A. Part 1 - Reproduction

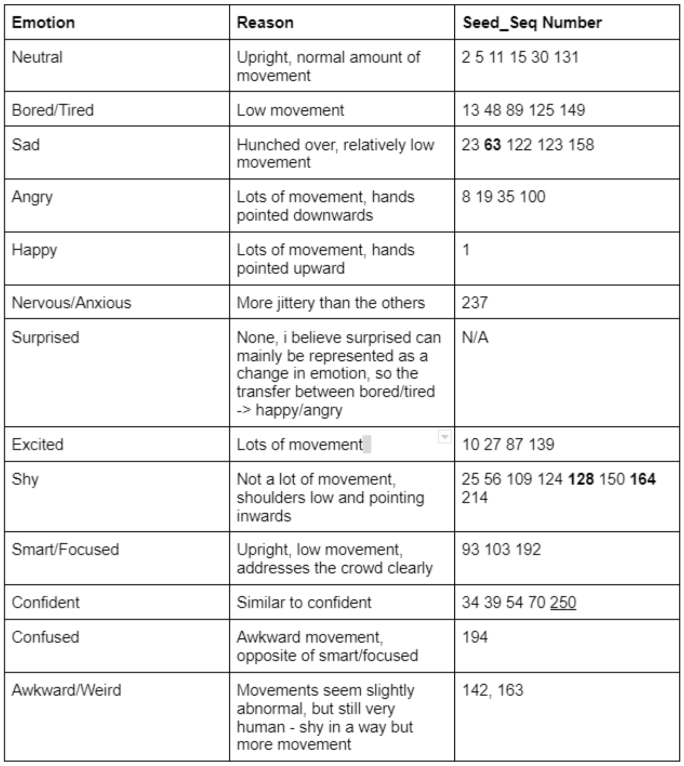
I will begin with the beginning of my semester. I first intended to work with Unity, but I had difficulties with utilizing an efficient graphics card. The vast majority of AI in VC requires this form of computation, so I turned to other resources. I attempted Google Colab, but was met with ceilings with storage and available processing. I then decided the best option would be the iLab. At this point, I switched focuses from Unity to the upper body gestures, as detailed previously. I began by implementing the speech2affectivegestures work. This was relatively difficult, since I needed to find access to the datasets. I was eventually able to find pretrained models which were much more efficient. From there, I exported everything to the iLab, where I had to work with the CS Dept to get access to larger amounts of user storage since it was not initially available to me as an undergrad. With regarding to the implementation, I also faced difficulties. There were file routing errors that had to be resolved, and several version errors. The program was produced a few years in the past, so I had to create a custom virtual environment on the iLab and update it to contain all of the necessary versions for the required libraries.

B. Part 2 - Modifications

Once I was able to get the program to run on my computer and reproduce the intended results, I worked on understanding what the inputs and outputs were, and how I would be able to edit them. For the first portion of my work, I wanted to be able to create the affect upper body gestures. In order to do this, I would first keep the text and the audio constant. From there, I would work on altering the seed poses and the seed sequences. I was able to change several portions of the work - changing some variables altered the current text/audio (would use this later). After several attempted modifications and errors, I was able to identify which parts of the code referred to the seed sequence which referred to the seed poses. At this point, I realized that changing the seed poses resulted in a very minimal difference, which was not nearly as pronounced or distinct as I would have wanted. The seed sequences on the other hand had the desired effect. I was able to save all of the seed sequences as a numpy array, and then load it back and iterate through it. As I iterated, I would keep all of the other variables constant, so there are less factors to consider. I was able to create 250 seed sequences on the iLab from this process, each using the same audio and text, but they each handled them in a slightly different way. They each changed the behavior of the set of upper body gestures and they seemed to represent a unique emotion.

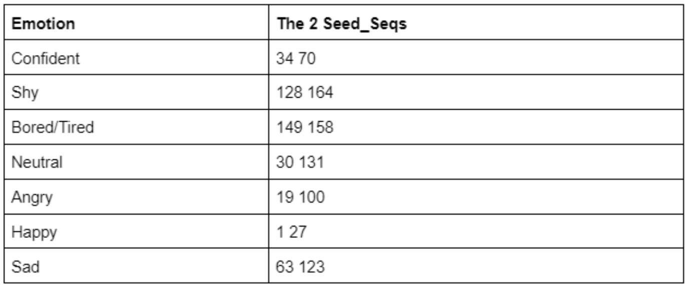
C. Part 3 - Emotion Analysis

I then watched all of the 250 videos with the varying seed sequences. My purpose here would be to determine which of these videos represent which emotion. The purpose of this would be to make emotion an input for this program instead of seed sequence, which would allow for the creation of affect gestures. My criteria and results for the initial 250 are shown below.



Emotion	Reason	Seed_Seq Number
Neutral	Upright, normal amount of movement	2 5 11 15 30 131
Bored/Tired	Low movement	13 48 89 125 149
Sad	Hunched over, relatively low movement	23 **63** 122 123 158
Angry	Lots of movement, hands pointed downwards	8 19 35 100
Happy	Lots of movement, hands pointed upward	1
Nervous/Anxious	More jittery than the others	237
Surprised	None, i believe surprised can mainly be represented as a change in emotion, so the transfer between bored/tired -> happy/angry	N/A
Excited	Lots of movement	10 27 87 139
Shy	Not a lot of movement, shoulders low and pointing inwards	25 56 109 124 **128** 150 **164** 214
Smart/Focused	Upright, low movement, addresses the crowd clearly	93 103 192
Confident	Similar to confident	34 39 54 70 250
Confused	Awkward movement, opposite of smart/focused	194
Awkward/Weird	Movements seem slightly abnormal, but still very human - shy in a way but more movement	142, 163

As you can see, not all of the 250 seed sequences are included, I focused on the ones that stood out to me, and were not repetitive. However, I ended up with too many emotions and corresponding seed sequences. So, I went through all of the previously selected ones to determine which were more accurate. I then arrived at the following:

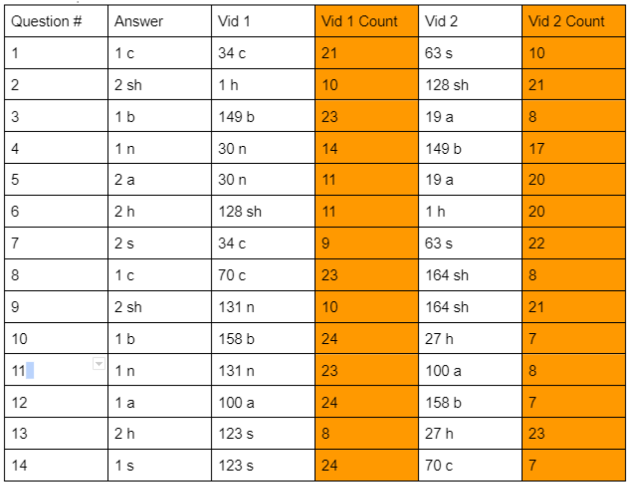


Emotion	The 2 Seed_Seqs
Confident	34 70
Shy	128 164
Bored/Tired	149 158
Neutral	30 131
Angry	19 100
Happy	1 27
Sad	63 123

These 7 emotions are all unique in their own way, and serve a purpose conversationally. I was also able to associate 2 seed sequences with each. Now, I had to design my statistical study to determine the accuracy of my initial assertions.

D. *Part 4 - The User Study*

From these 7 emotions and 14 seed sequences, I designed and carried out a user study. This user study consisted of 14 questions, each asking users to choose between two videos. The videos all contained the same text and audio, and the only difference was the seed sequence. Each question would ask, for example, "Q1 - 1/14 Which set of upper body gestures better represents confidence?". One of the answer choices would be a seed sequence I deemed to represent confidence, and the other would be some random seed sequence from the remaining 13. With 14 questions, each seed sequence was a correct answer for one question and an incorrect answer for another question. The results of the study are shown below.



Question #	Answer	Vid 1	Vid 1 Count	Vid 2	Vid 2 Count
1	1 c	34 c	21	63 s	10
2	2 sh	1 h	10	128 sh	21
3	1 b	149 b	23	19 a	8
4	1 n	30 n	14	149 b	17
5	2 a	30 n	11	19 a	20
6	2 h	128 sh	11	1 h	20
7	2 s	34 c	9	63 s	22
8	1 c	70 c	23	164 sh	8
9	2 sh	131 n	10	164 sh	21
10	1 b	158 b	24	27 h	7
11	1 n	131 n	23	100 a	8
12	1 a	100 a	24	158 b	7
13	2 h	123 s	8	27 h	23
14	1 s	123 s	24	70 c	7

Active Link to the study

E. *Part 5 - Analysis*

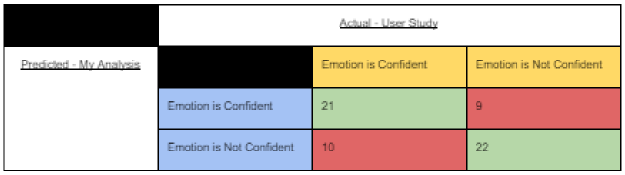
From this design, I was able to create a confusion matrices for each of the 14 seed sequences. My assertion would be the method, and I would consider the user study to be the truth. When a given seed sequence is the correct answer and the user chooses the correct seed sequence, that would be a true positive, when the other choice is chosen that would be a false positive. For when the given seed sequence is incorrect and it is chosen that would be a false negative, and when the other option is chosen that would be a true negative. From this, I was able to export all of the data from the study into these confusion matrices. A full list of these is available at the following link:

Link to all of the Confusion Matrices.

A few examples are shown on the right as well

Confusion Matrices for Confidence:

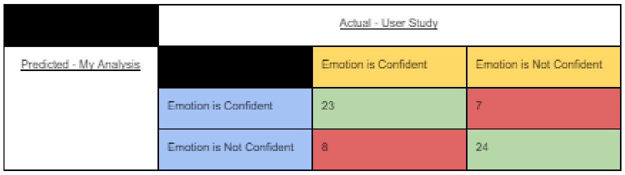
34



	Actual - User Study		
Predicted - My Analysis		Emotion is Confident	Emotion is Not Confident
	Emotion is Confident	21	9
	Emotion is Not Confident	10	22

Sensitivity: 67.74
Specificity: 70.97
Accuracy: 69.35

70

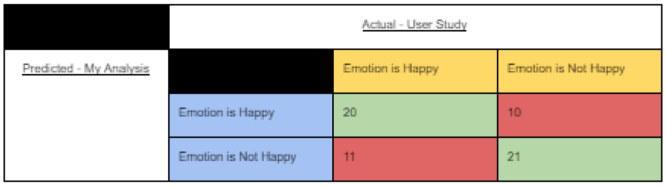


	Actual - User Study		
Predicted - My Analysis		Emotion is Confident	Emotion is Not Confident
	Emotion is Confident	23	7
	Emotion is Not Confident	8	24

Sensitivity: 74.2
Specificity: 77.42
Accuracy: 75.8

Confusion Matrices for Happiness:

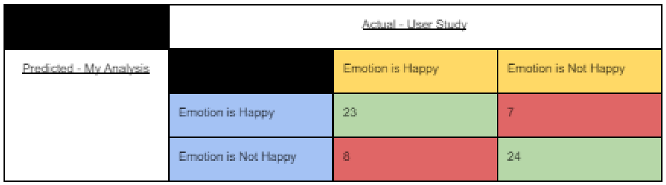
1



	Actual - User Study		
Predicted - My Analysis		Emotion is Happy	Emotion is Not Happy
	Emotion is Happy	20	10
	Emotion is Not Happy	11	21

Sensitivity: 64.52
Specificity: 67.74
Accuracy: 66.13

27

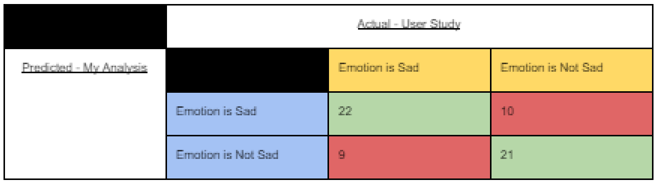


	Actual - User Study		
Predicted - My Analysis		Emotion is Happy	Emotion is Not Happy
	Emotion is Happy	23	7
	Emotion is Not Happy	8	24

Sensitivity: 74.2
Specificity: 77.42
Accuracy: 75.8

Confusion Matrices for Sadness:

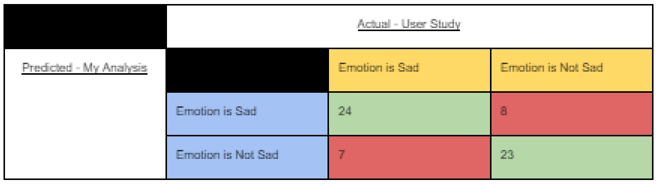
63



	Actual - User Study		
Predicted - My Analysis		Emotion is Sad	Emotion is Not Sad
	Emotion is Sad	22	10
	Emotion is Not Sad	9	21

Sensitivity: 70.97
Specificity: 67.74
Accuracy: 69.35

123



	Actual - User Study		
Predicted - My Analysis		Emotion is Sad	Emotion is Not Sad
	Emotion is Sad	24	8
	Emotion is Not Sad	7	23

Sensitivity: 77.42
Specificity: 74.2
Accuracy: 75.8

Overall the data was relatively accurate. It is reasonable to assume that there is no bias with regard to the people taking the survey, there is essentially no correlation between race or age or gender and emotion recognition classification. All portions of the setup of the study were random as well. The only potential source of bias is that this study only required the users to decide between two emotions. For comparisons between angry and boredom for example, the results were much stronger than of those between sadness and boredom. This is because some emotions are more/less associated with other emotions. In order to avoid this, one could instead provide several more questions and increase the sample size, but that is relatively difficult to do. However, from this user study and the available data, I am now able to create an upper body gesture with some given text and audio that can represent an emotion of the user's choice, therefore asserting a correlation between emotion and upper body gestures.

F. Part 6 - Text to Speech

Previously, I was able to implement Nvidia Tacotron2, which is an available TTS. This program however was running off of the iLab, which made integration difficult. So, I decided to work with Google Cloud's TTS instead, which was implemented in python similar to the STT, NLU RASA, and the modified upper body gestures. I was able to implement it with relative ease, but it was not affect, it simply provided several options for language/accent/gender. I again continued to search for a more suitable TTS. I attempted to implement that of typecast.AI, but they all either required subscriptions or provided a limited number of customization options which was not ideal. Regardless, this is something that is independent of the upper body gestures. It only requires audio, text, and an emotion to compute the desired set of gestures. So, I then looked into altering the program so it can accept custom text and custom audio.

G. Part 7 - Custom Text and Audio

Given additonal time, I have begun working on incorporating custom text to speech. There are several different components to modify this. I realized that there is some number that if I were to modify, provides me with a different text/audio. So, I wanted to see where that number was used, and it was a total of 4 places, in skeletons3d, audio, starttime, endtime, and words. These are all data structures that only correspond to this given audio/text number. I am able to modify starttime (a double), endtime (a double), words (an array of each word being said and its correlating timestamp), and audio (the audio file itself). Skeletons3d seems to be a ground truth, so I have no decided to leave it constant. I have modified the other variables, but I am currently running into other errors that require debugging. I have been working on this for the past few days, and will continue this work

V. Conclusion and Future Work

Overall, I was able to complete the intended work and arrive at telling results. We now have data supporting which seed sequences represent which emotion, which add to the humanity of the embodied conversational agent. I was also able to implement the speech2affectivegestures project in python. Now, we can work to create an overall more complete product. I begin by speaking to the computer, where my speech gets converted into text. That text then gets mapped to some intent, and some response text and emotion are computed. From there, we can create affect audio given the text and the emotion. Using these three, we can create the affect upper body gestures, and several other portions of the ECA such as facial reconstruction. Looking forward I believe the team will be able to accomplish this under Jerry, working to create the fully functional and accurate ECA.

VI. Acknowledgement

This work is part of an ongoing research project at Rutgers University under Professor Mubassir and Jerry Chang.

References

[1] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2027–2036, 2021.

[2] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1359–1367, 2020.

[3] Sinan Sonlu, Uğur Güdükbay, and Funda Durupinar. A conversational agent framework with multi-modal personality expression. *ACM Transactions on Graphics (TOG)*, 40(1):1–16, 2021.