Gesture Generation and Embodied Conversational Agents

Abdullah Ghanima, Jerry Chang, Mubbasir Kapadia

Department of Computer Science, Rutgers University, Piscataway, New Jersey 08854

Aresty Research Center for Undergraduates

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, we are 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). For this past semester I have been working on the graphical components, synthesizing audio (tone/word choice) of a user with their upper body gestures, with the purpose of better simulating the virtual human. In order to assess the degree to which we are accurately representing human emotion, I have carried out a statistical study as well. This is a continuing work, and we hope to place it into effect in the near future.

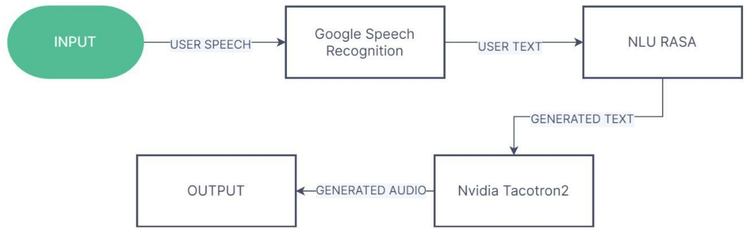
Background

Arguably, one of the best ways to improve economic mobility in underprivileged neighborhoods is to improve the quality of the teachers. We hope to create a virtual training tool that has the purpose to better equip these teachers to handle students and any correlating difficult situation. 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. This project is extensive and contains several parts.

We are also 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.

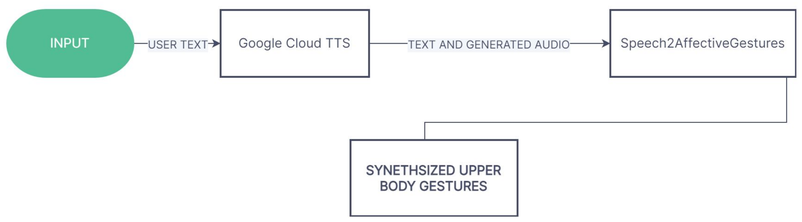
Materials and Methods

I began my research working on creating a basic conversational AI. This tool contained several technologies that were connected via python programming. First, the user inputs some audio in the form of speech. That audio gets processed by Google Speech Recognition, which converts it into the spoken text. That text gets run through a natural language understanding named RASA, which runs on a local server. After manipulating RASA and creating a few sample hypothetical scenarios, some output text would be computed which would then get sent to Nvidia Tactotron2, which is a text to speech program running off of Google Colab. This would output the text as speech. This concluded the basic conversational AI, which receives user input speech, and outputs a corresponding response.

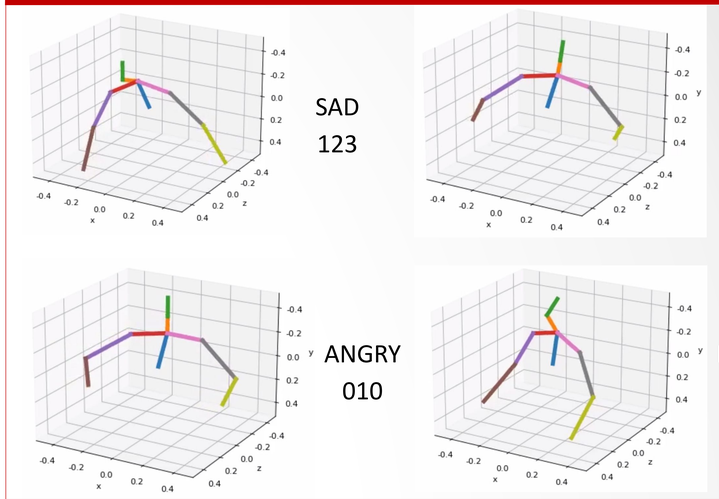


I then reproduced the work done in Speech2AffectiveGestures which is a project that creates upper body gestures through neural networks. It uses the Ted and the Trinity datasets to synthesize a relationship between upper body gestures and input text, input audio, user ID, and seed poses. I first isolated seed poses as a variable, and generated 172 unique gestures using the same audio/text. I have determined which of these represent which emotion, and I will carry out a statistical study in coming weeks to assess its validity.

I have also implemented Google Cloud Text to Speech, an affect TTS. I will use this in conjunction with the modified Speech2AffectiveGestures work in order to create simulated upper body gestures with associated audio and emotion.



Results



Conclusion

This academic year has yielded substantial progress, we were able to create a basic conversational agent, and implement upper body gestures. We have also found relationships between the emotion of the agent, and its speech/gestures.

Future Direction

I will be continuing my research for the remainder of the semester through my work in CS 549. I will first complete the statistical study, to prove the validity of the correspondence between emotions and upper body gestures. After that, I will connect the Google Cloud TTS to the Speech2AffectiveGestures work and write a formal research paper about m work.

For the far future, the work conducted in this semester will become a part of the finalized embodied conversational agent. Work will be done with psychologists and teaching professionals in order to determine agent specifics, and the product will then be placed into practice.

Acknowledgments

- Speech2AffectiveGestures: Synthesizing Co-Speech Gestures with Generative Adversarial Affective Expression Learning - *Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, Dinesh Manocha*
- Professor Kapadia and Rutgers CS 549 - Artificial Intelligence in Visual Computing
- Jerry Chang and the the ECA team