Robot Learning Project Proposal

Anna Buchele, Ariana Olson

- In 4-5 sentences, what is the big idea of your project?
    - Our project is to train a robot to respond to a person's vocal emotional expressions. For example, when a person yells at the robot, it might run away. But if the person speaks kindly to the robot, it might approach. This will require a few things: a large set of both positive and negative recordings, localization of the source of the sound, and basic navigation elements such as obstacle detection and target approach.
- Have you found any papers or blog posts that have done something similar to what you are proposing?
    - https://cacm.acm.org/magazines/2018/5/227191-speech-emotion-recognition/fulltext This is a post describing traditional ways to characterize emotion from acoustic characteristics of a speech sample
    - Localization using time difference of arrival: https://pdfs.semanticscholar.org/61b9/ab86eecbeb419f5bbe1da6bc5e6372c3c9ad.pdf
    - Some papers on using NNs for emotion recognition:
        - https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=845644
        - http://www.ccis.northeastern.edu/home/daikeshi/papers/iasted08.pdf
    - Someone's implementation:
        - https://github.com/MarioRuggieri/Emotion-Recognition-from-Speech
    - Hidden Markov model: https://www.sciencedirect.com/science/article/pii/S0167639303000992?via%3Dihub
- In terms of the system that you will have running on the robot, what is your MVP? What is your stretch goal?
    - MVP: robot responds to emotional expression present in a single word, blindly approaching the source for positive, and blindly running away for negative.
    - Stretch goal: robot responds to emotional expression for a variety of words and phrases, and smartly (using obstacle detection, etc) approaches or runs away based on the emotions present.
- Describe your learning orientation (top-down versus bottom-up) and why you have chosen it. In particular if you choose bottom-up, make sure you specify what this will mean (e.g., which algorithms will you implement, will you eventually switch to a standard toolkit, etc.).
    - Our learning orientation will be bottom-up, because we haven't found any out-of-the-box solutions for this sort of problem, and we would like to learn more about how machine learning works. We plan on using an existing toolkit (PyAudioAnalysis) to extract the features of the sound, and then creating our own

algorithm to classify those features as negative/aggressive or positive/friendly. We have found a few github projects where people did similar things to reference if we get lost.

- What is your data collection plan? How do you plan to get the data needed for your project? How much data do you think you'll need? Are there existing datasets you can leverage?
    - We plan on finding some datasets online, but also recording us and our friends saying things nicely and rudely. More specifically, we will choose a set of phrases and have friends say them in both “nice” and “mean”
- What sorts of learning algorithms will you apply? You could choose these based on what you think will work the best or what you want to learn about the most.
    - We are planning on using a deep neural network or a convolution neural network to classify the emotions in the same way an image would be classified
        - We are still deciding on the specific architecture to use based on papers we are looking at. One approach would be to use a DNN or CNN to create a single classifier network. Another interesting approach described in one of the papers linked above would be to use two separate networks that can be tuned individually to determine the likelihood of each emotion.
- What non-learning baseline algorithm will you compare to?
    - A more “classical” approach would be to use a hidden Markov model to perform emotional analysis: https://www.sciencedirect.com/science/article/pii/S0167639303000992?via%3Dihub
    - A simpler non-learning approach would be to find several features in our data that point to one emotion being more likely (such as pitch, volume, etc.) and using these to characterize the audio.
- Also, in order for this to succeed we would need three microphones for the localization.
    - (Can we?)