AI Companion:
Speech, Emotion and Context Recognition

Prutha Parmar (prp2126) and Liane Young (ly2451)

Topic A2: Speech Recognition, Gesture Recognition, and Feeling Recognition

February 14, 2020

Abstract: While there are many examples of virtual companions such as Siri and Alexa that are available on the market today, they often lack empathy in their ability to respond. This project proposes a solution to provide a supportive companion that is able to consider the emotional needs of the user. Through the use of speech emotion recognition and a context dialogue model, we will be able to formulate appropriate responses that can read the mood of the conversation and respond in a logical manner.

1. Background (Review of Related Literature):

In today's world, we are aware that there are alot of conversational agents such as Siri, Google Home, Cortana, Alexa, etc. While these dialogue bots are deemed to be very successful, they have their own shortcomings. For instance, after doing a lot of research it has been found that these conversational agents are insensitive and fall short of responses when it comes to emotional situations. This problem is because of the limited vocabulary that they have to choose from and come up with responses. A lot of research work is going on in this domain where conversational agents are trained to help people in mental health situations. For example, chatbots like Woebot, 7Cups and Koko have been used for various tasks, such as providing psychological assessments [2]. Therefore, we are trying to overcome this challenge of creating an emotional agent which at the very least can connect a person to help.

There has been many studies into speech emotion recognition (SER) systems. Many SER systems that have been studied use statistical pattern recognition techniques to perform classification based on their emotional content [3]. Over the past several years, the rise in popularity of machine learning techniques and neural networks have given way to systems that can now classify emotion in speech with up to 84% accuracy in noisy conditions [4]. The research in these prior studies present techniques that can be used to improve upon traditional dialogue bots as we attempt to make them more empathetic.

2. Introduction to the Project:

In order to achieve this project's goals, our system will require several methods including data pre-processing, emotion recognition, speech recognition, and developing a context dialogue model. An overview of the proposed system can be found in Figure 1.

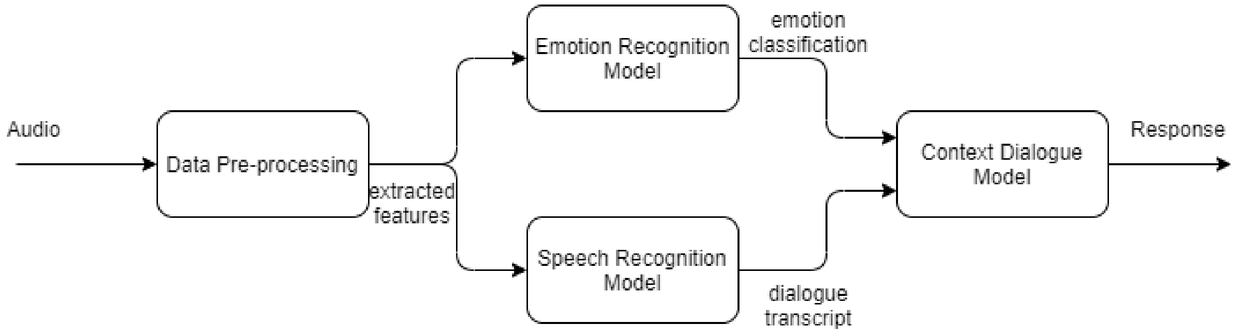

Figure 1: Proposed system architecture for emotion contextual bot

This project will require data pre-processing in order to prepare the system inputs for speech and emotion recognition. First, samples of the audio input will be taken for some window length. The Mel spectrogram of each of these samples will be obtained as a frequency-based representation of the data. Feature extraction can then be performed on the spectrograms to determine what to feed into the speech and emotion recognition models.

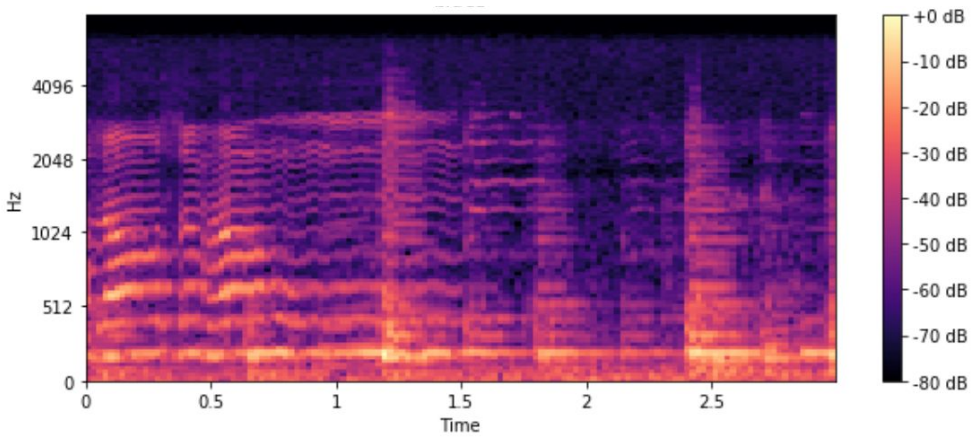

Figure 2: Example of Mel spectrogram that will be used to perform feature analysis

Three different machine learning models will be required. For each type of model needed, we will be performing experiments to determine the best performance we can achieve for each component.

Emotion recognition poses a classification problem, so we perform experiments to determine the best model for characterizing emotions in speech. We will weigh the results of a basic artificial neural network (ANN), K-nearest neighbors algorithm, and a Gaussian model to determine the best results.

For speech recognition, we need to determine how to best get a speech audio file converted to text so that it can be used in the next steps of the system. For this model, we plan to investigate using a Hidden Markov model or recurrent neural networks (RNN).

Our dialogue model requires a model that contains a memory system that can remember previous inputs for context, and determine the best responses of both the current inputs and prior context. Based on our literature survey, we want to use the Seq2Seq model as proposed by Wei et al. [1] but if results aren't good, we may perform experiments on other machine learning models that have memory capabilities.

Ultimately, the system will be able to produce a response based on the emotion of the speaker and contextual clues of the conversation.

3. Introduction to the Dataset:

This project will require data to train both the speech emotion recognition model and the context dialogue agent.

For the speech emotion recognition model, we have identified three viable datasets. Our primary dataset we plan to train with is the RAVDESS dataset, which is a public dataset that consists of 7356 audio files using the English language. RAVDESS contains 8 different emotions (neutral, calm, happy, sad, angry, fearful, disgust, and surprise) at two different emotional intensities, and both speech and song audio files are available. The other two potential datasets are the CREMA-D dataset and the MSP-IMPROV, which are less emotionally diverse datasets. However, one issue common when training speech emotion recognition models is gathering a large enough dataset to properly train the model. If we encounter this issue while training with the RAVDESS dataset, we plan to either use data augmentation methods such as introducing noise into the dataset or incorporating the less emotionally diverse data in the training process as well.

We have also identified several different dialogue datasets that can be used to train the model to generate responses. The top 3 datasets we have found are the Ubuntu Dialogue Corpus, the Santa Barbara Corpus of Spoken American English, and the Cornell Movie Dialogue Corpus. The Ubuntu Dialogue corpus includes 930,000 dialogues about different issues related to Ubuntu. The Santa Barbara Corpus includes 249,000 words of conversations on different topics, with both spoken dialogues and their transcripts provided. The Cornell Movie Dialogue Corpus contains 220,579 exchanges from movies, including the dialogue and their scripts. All of these datasets are available to the public. The primary obstacle we will need to deal with for the dialogue datasets will be gathering diverse enough data to train a model well. The model needs a large vocabulary in order to get context correct in execution, but needs to be trained on a large and diverse lexicon in order to gain breadth in its vocabulary.

4. Plan:

Milestone 1

For the first milestone, we performed a literature survey. We studied the various limitations that the previous studies posed such as the use of statistical pattern recognition techniques, limited set of vocabulary which leads to failure in understanding the meaning and thus, failure to generate a proper response. On the other hand, we also studied the recent advances in this field about how supervised or semi-supervised machine learning techniques are able to correctly predict the emotions in noisy environments.

This helped us formulate our plan to tackle this challenge and helped us decide what methods we should deploy.

Milestone 2

For the second milestone, we wish to develop various machine learning models such as artificial neural network, KNN and gaussian model for emotion recognition task. For speech recognition, we plan on using hidden markov model or recurrent neural network. Moreover, for our dialogue agent model we want to use a sequential model since it has given promising results.

Milestone 3

For the third milestone, we want to train our models and generate the response. The plan is to compare these models and select the one which gives the best accuracy for each of the above discussed tasks. For comparison, we will be looking at various evaluation metrics such as BLEU, word accuracy and emotion accuracy. We also aim to develop an interface for the system.

Final Project

After all the milestones are completed, for our final project we want to collaborate with Team A1 with the following group members: Raksha Ramesh and Kavita Anant. We wish to develop an interface that integrates both our projects i.e. face and speech generated emotion models. Moreover, we want to refine our project by adding context and dialogue features to it.

Reference:

1. H Wei, et al. “Building Chatbot with Emotions”. http://web.stanford.edu/class/cs224s/reports/Honghao_Wei.pdf
2. Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions https://www.jmir.org/2018/6/e10148/

3. F. Dellaert et al. “Recognizing emotion in speech”. 1996. https://ieeexplore.ieee.org/document/608022
4. Mehmet B. Akcay and Kaya Oguz. “Speech emotion recogntion: Emotional models, databases, features, preprocessing methods, supporting modlities, and classifiers”. January 2020. https://www.sciencedirect.com/science/article/pii/S0167639319302262