Yun Jia Zhuang & Janelle Tam Feb 8th, 2024

MAIS 202 Project Proposal: Emotion Classification from Audio Data

1. Choice of dataset:

Here are the datasets we will be using:

- Surrey Audio-Visual Expressed Emotion (SAVEE)
  - Male actors expressing seven different emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral.
  - Good data quality: not too much background noise, audio files are about 3 seconds long (good for single, consistent emotion)
  - Distinct differences in audio waves from one audio file (e.g. happy) to another
- RAVDESS Emotional speech audio
  - Emotions: neutral, calm, happy, sad, angry, fearful, disgust, surprised.
  - 'Emotional intensity' feature (normal, strong), except for 'neutral' emotion.
  - Only audio files will be used.
- Toronto emotional speech set (TESS)
  - Female actors expressing seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral).
- CREMA-D
  - Male and female actors, wide range of age, across different ethnicities
  - Six emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad)
  - Four emotion levels (Low, Medium, High, and Unspecified).
- IEMOCAP- Home
  - Emotions: anger, happiness, sadness, neutrality
- EmoReact Dataset
  - Audio files of children (ages 4-14)
  - 17 different emotional states

****Note:** We do not yet have access to IEMOCAP and EmoReact datasets; we have requested these datasets from their holders.

To prevent overfitting, we want to use datasets with various speakers in different environments, so our model can perform well even with unfamiliar audio input.

2. Methodology:

High-level overview: We want to train a model to automatically classify the emotion of a given piece of audio data. As a secondary component, we want to then use this model to perform a novel task (e.g. generate a scene, play a song, find a quote or book passage).

a. Data Preprocessing:

1. We will explore each dataset to determine the optimal method to mix and match the files. We also want to make sure that the audio files are distinguishable enough from each other that our model will be able to classify the emotion
   a. e.g. Group all the files across multiple datasets labelled with the same emotion together
2. We will mainly consider the "emotion" label and disregard the "intensity" label, as the latter is not present across all datasets.
3. We will also need to convert our audio files into a format that our model can process

   a. Specifically, we will be using LibROSA to convert input .wav files into an MFCC image format.
   b. We will consider the pitch, frequency, energy, and waveform features, among others.

4. We will normalize the data due to our multiple datasets.
   a. From this paper, Z-score normalization is shown to work well with CNNs.

b. Machine learning model:
We want to predict emotion from a given audio file.
Below are some models that can accomplish this:

1. **Random Forests**
   a. Pros: Multiple decision trees result in a lower likelihood of one incorrect decision tree classification impacting the overall result
      i. Can handle different types of audio files well, even if some of the audio clips are noisy
   b. Cons: More computationally intensive due to multiple decision trees
      i. Lack of interpretability; “black box”
2. **Support Vector Machines (SVM)**
   a. Pros: Good at finding clear boundaries between emotions
   b. Cons: Can be slow with large datasets; may be confused with data where there are overlapping emotions
3. **CNNs**
   a. **Pros:** Good at detecting patterns in spatial data, such as the spectrograms of audio signals
      i. Learns on its own from examples
   b. Cons: Requires a large amount of labeled data
      i. Lack of interpretability; “black box”
4. **Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)**
   a. Pros: Good for data where the order is important, such as the case where a changing tone in an audio clip indicates a specific emotion
      i. Can reference earlier parts of the clip to make a more holistic guess
   b. Cons: Complex to use; model may struggle with longer audio clips

According to multiple sources, including this paper and this article, CNNs and RNNs seem to be the most popular for the types of audio classification problems we want to solve. We want to test each of these models, however, to see which one gives us the highest accuracy.

c. Evaluation Metric:
Since we are working on a classification problem, we will be using confusion matrices to evaluate the accuracy and precision of our model.

3. Application:
We will be looking into creating a webapp, where users will be able to record an audio clip and submit it for our model to assess emotion. As output, the user will receive a list of top 3 emotions predicted by the model, in decreasing order of probability. As mentioned above, if we have time, we want to then use this model to perform a novel task (e.g. generate a scene, play a song, or find a quote or book passage).