Speech Emotion Recognition

Team Members:

Punalraj Parthasarathy (A20519421) pparthasarathy@hawk.iit.edu

Santhosh Mani (A20518627) sm3@hawk.iit.edu

Abstract:

Emotion recognition helps in monitoring and understanding human emotion, which is important for both existing and upcoming computational technologies. Speech emotion recognition is a demonstration of the ability to recognize emotional states and human feelings from speech. The experiments take neutral, happy, and sad into consideration. Physical traits including muscle tension, skin elasticity, blood pressure, heart rate, breathing, speech, etc. have a major impact on a person's emotions. Although each person's emotions are distinct in nature, they can nevertheless be understood, perceived, and reflected in different ways. Python libraries are used to do this investigation.

Problem Statement:

In this project, we'll develop a model using an MLPClassifier using the libraries librosa, soundfile, and sklearn (among others). This machine learning model will be able to identify emotions from audio files. After loading the data and extracting its features, the dataset will be divided into training and testing sets. The model will then be trained after initializing an MLPClassifier. We will then determine the model's accuracy.

Methodology:

- As we are working with two different datasets. So, we will be creating a data frame storing all emotions of the data in a data frame with their paths.
- We will use this data frame to extract features for our model training.

- **Data Augmentation:**
  - Data augmentation is the process by which we create new synthetic data samples by adding small perturbations on our initial training set.
  - To generate syntactic data for audio, we can apply noise injection, shifting time, changing pitch and speed.
  - The objective is to make our model invariant to those perturbations and enhance its ability to generalize.
  - In order for this to work adding the perturbations must conserve the same label as the original training sample.
  - Just like In images data augmentation can be performed by shifting the image, zooming and rotating.

- **Feature Extraction**
  - Extraction of features is a very important part in analyzing and finding relations between different things. As we already know that the data provided of audio cannot be understood by the models directly so we need to convert them into an understandable format for which feature extraction is used.
  - The audio signal is a three-dimensional signal in which three axes represent time, amplitude, and frequency. With the help of the sample rate and the sample data, we can perform several transformations on it to extract valuable features from it. Some of the features that could be extracted are listed below :
    - Zero Crossing Rate: The rate of sign-changes of the signal during the duration of a particular frame.
    - Energy: The sum of squares of the signal values, normalized by the respective frame length.
    - Entropy of Energy: The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.

    - Spectral Centroid: The center of gravity of the spectrum.
    - Spectral Spread: The second central moment of the spectrum.
    - Spectral Entropy: Entropy of the normalized spectral energies for a set of sub-frames.
    - Spectral Flux: The squared difference between the normalized magnitudes of the spectra of the two successive frames.
    - Spectral Rolloff: The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
    - MFCC: Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
    - Chroma Vector: A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
    - Chroma Deviation: The standard deviation of the 12 chroma coefficients.
  - In this project, we are only extracting 5 features out of the few mentioned above:
    - Zero Crossing Rate
    - Chroma_stft
    - MFCC
    - RMS(root mean square) value
    - Mel Spectrogram to train our model

- **Data Preparation**
  - As of now we have extracted the data, now we need to normalize and split our data for training and testing.
  - Then we will be initializing an MLPClassifier which is a Multi-layer Perceptron Classifier; it optimizes the log-loss function using stochastic gradient descent. Unlike Support Vector Machines (SVMs) or Naive Bayes, the MLPClassifier has an internal neural network for the purpose of classification. This is a Feed Forward Neural Network model.

- **Modeling**
    - We will train the model using the train data to make it predict the emotions of the respective audio files.
    - We will use the test data to predict the emotion of the test data.
- Then we find the accuracy of our trained model using the Sklearn library by comparing the predicted and target values.

Description of the dataset:

Datasets used in this project:

- Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess)
- Crowd-sourced Emotional Multimodal Actors Dataset (Crema-D)

Dataset 1: (Ravedess)(Link)

This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions include calm, happy, sad, angry, fearful, surprised, and disgusted expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

File naming convention

Each of the 1440 files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav). These identifiers define the stimulus characteristics:

Filename identifiers

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

Filename example: 03-01-06-01-02-01-12.wav

1. Audio-only (03)
2. Speech (01)
3. Fearful (06)
4. Normal intensity (01)
5. Statement "dogs" (02)
6. 1st Repetition (01)
7. 12th Actor (12)
   Female, as the actor ID number is even.

Dataset 2: (Crema-D)(Link)

CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

Workdone so far:

- We have Trimmed the datasets to match our project needs.
- We have completed data preprocessing.
  - ➔ Made sure that we did not have any duplicate files.
  - ➔ Removed any outliers.
- Converted the datasets into data frame with respective file paths.
- We have applied data augmentation and extracted the features for each audio file and saved them.

Future Works

- We will be normalizing and splitting our data for Training and Testing.
- We will be initializing the MLP classifier.
- We will use the Test data to predict the emotions.
- Then Using the Sklearn Library we find the accuracy of the trained model by comparing the predicted and target values.