SpeakFeel: Emotion Recognition In Speech

Team Members

Sike Ogieva, Kien Tran, and Sauryanshu Khanal

Purpose and Vision

Emotion recognition has important applications in improving NLP; in psychology, customer service, human computer interaction and emotional AI. It is a growing field, and one that is crucial for teaching technology about humanity and fueling the ongoing transformation of user interfaces.

The aim of this project is to build a robust, accurate neural network model capable of recognizing and classifying a range of human emotions from speech data. Given that this is a task even humans can not always reliably do, we are excited to see what our machine attempt looks like. Our goal is to bridge the gap between human emotional expression and machine interpretation, enhancing human-computer interaction across various applications.

Data

We will be using this Kaggle dataset (the audio-speech subset of the Ryaerson Audio-Visual Database of Emotional Speech and Song) . The dataset comprises WAV files, with each 24 actors contributing 60 recordings, totaling 1,440 files. Actors are evenly divided by gender, with men assigned odd numbers and women even numbers. Every actor delivers their lines in a level North American accent. Files are identified by emotion (neutral, calm, happy, sad, angry, fearful, disgusted and surprised); emotional intensity (normal and strong), the text of their statement, and the speaking actor. Neutral emotions have no intensity.

Given that our task is a basic classification problem, 1440 files with a wide variety of emotions prove to be of sufficient data for training our neural network. We are open to sourcing (and creating) additional data if needed.

Process

Audio feature extraction can be done by using the Python library “librosa” to farm Mel-Frequency Cepstral Coefficients (MFCCs) and wavelet transform features which both mimic human hearing, or creating spectrograms which neural networks can take in as an image. Other options are to exploit the inherent features of convolutional neural networks that let them automatically learn to extract relevant features; or to use autoencoders.

We'll start with preprocessing the audio to normalize and extract features like MFCCs, followed by an exploratory data analysis to understand the emotional distribution and actor balance

We will select hyperparameters for our neural network by referencing literature, and with experimentation. In addition to the neural network, we will implement two baselines (kNN and human classification). Our model's performance will be meticulously evaluated using accuracy, precision, recall, and F1 score, with a focus on hyperparameter tuning and data augmentation to ensure robustness and generalizability.