Erdős Institute Bootcamp May 2022
Team Genesis

Executive Summary

Classifying Emotion from Human Speech

Correctly classifying human speech into categories of emotions, based solely on inherent characteristics of the audio (i.e. spectral features), has an ever-growing importance in many regards. From the more general goal of integrating human emotions into AI technology to **helping the millions of autistic individuals perceive the subtleties** that are being communicated to them or the multi-million dollar **audio translation and speech-to-text market**, the list of stakeholders in an accurate emotion classification goes on.

However, research shows that visually and audio-visually communicated emotions are more readily perceived and more accurately classified by human raters than audio-only ones. Following the recent strides made in Machine Learning techniques, we ask the questions: **how do we leverage the power of Machine Learning to improve emotion classification of audios?** To answer this question, we used the Crowd-sourced Emotional Multimodal Actor Dataset (CREMA-D). The dataset is collected from 2,443 human raters who were tasked with classifying a set of audios (in addition to visual-only and audio-visual) into **six emotion categories, namely Happy, Anger, Sad, Fear, Disgust and Neutral**. The audios were recordings of 12 sentences, each produced with 6 emotions by 91 actors from a wide age range and across different cultural backgrounds.

The input to our classifiers are 6,076 actor-produced audio recordings (the data). Only 11 of the 12 sentences were used, as 1 of them included a variable that we ignored in the present project. We pre-processed the data by reducing the unwanted noise and removing the leading and trailing silences with audio-analysis packages. From here, we took two approaches. On the one hand, we **extracted spectral features** (i.e. mid-term features) from the audio and, on the other, we **created spectrograms**, which are images that represent an audio signal. These representations allowed us to approach the classification problem in two different ways: both as an audio classification problem, and an image classification one.

With these two types of representations, we built two machine learning pipelines, each using a different data representation. We used a Support Vector Machine (SVM) on the spectral features and a Convolutional Neural Network (CNN) on the spectrograms. After several parameter tunings on both techniques, the classification accuracy was 42% for the CNN and 45% for the SVM against 40.9% for human raters. The SVM model is found to be more promising to detect the Neutral and Anger emotion categories, while CNN worked best at detecting Sad and Anger. Crucially, when the classification is done over a subset of the six emotions, **the accuracy of the SVM drastically increases, going as high as 90% when classifying Disgust vs Fear**. The low accuracy observed when classifying all 6 emotions may be due to incorrect labeling because the target variable was the intended emotion and not the one actually produced by the actors.

In summary, our classifiers performed at least as well as human raters did, and beyond. The next step in our project is to use much cleaner data, such as one that is not staged by actors like the CREMA-D (e.g: phone conversations) or maybe a subset of the CREMA-D data whose labels match the classification of the human raters. With such data, the classifiers may perform more accurately, to the delight of our many stakeholders.