Emotion Recognition Using Speech

Jacob Fenger

Motivation and Background

- **Rise of devices that utilize speech in recent years**
- **Machine detection of emotion would enhance user interaction**
- **Problem: Emotion is inherently complex**

Real Life Applications:

- Responses from Amazon Alexa are altered based upon the speaker's emotional state (Adaptability)

Overview

Goal: Accurately classify emotion based upon speech signals

Steps:

1. Obtain human speech signal data

1. Perform feature extraction on the data

1. Train a classifier for emotion detection based upon extracted features

The Data

Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database

- Contains 12 hours of audiovisual data of actors performing improvisation and scripted scenarios
- Audio dialog was segmented at the dialog turn level (A continuous segment of an actor speaking)
- 10039 turns in total
  - 5225 scripted
  - 4784 improvisational
  - Average duration of 4.5 seconds
- Also contains emotional labels for each turn
- More info: http://sail.usc.edu/iemocap/

What I did:

- Only utilized the improvisational turns
- Performed feature extraction on each of the turns using the openSMILE tool
- Parsed emotional labels to assign categories to each sample
  - Happiness, Sadness, Frustration, Anger, Surprise, Excited, Fear, Disgust, Other

Feature Extraction

openSMILE:

- A free tool to extract audio features from sound files
- Able to specify config file (Which dictates what features are extracted) and output
- Example usage:
  - *./SMILExtract -C config/IS09_emotion.conf -I example.wav -O output.csv*

Features Extracted:

- Utilized the INTERSPEECH 2009 Emotion Challenge feature set
  - This contains 384 features as ‘statistical functionals applied to low-level descriptor contours’
    - Root-mean-square signal frame energy
    - Zero-crossing rate of time signal
    - Mel-Frequency cepstral coefficients
    - And several more...

Classification

- **Gaussian Mixture Model**
  - Utilized Scikit-learn to generate the model
  - Number of components found via 5-fold cross validation

- **Classification accuracy with 5 components on test set:**
  - **16.823%**
  - Not good.

- **Problem:**
  - There were too many classes with an unevenly distributed data set
  - Some samples had multiple classes assigned

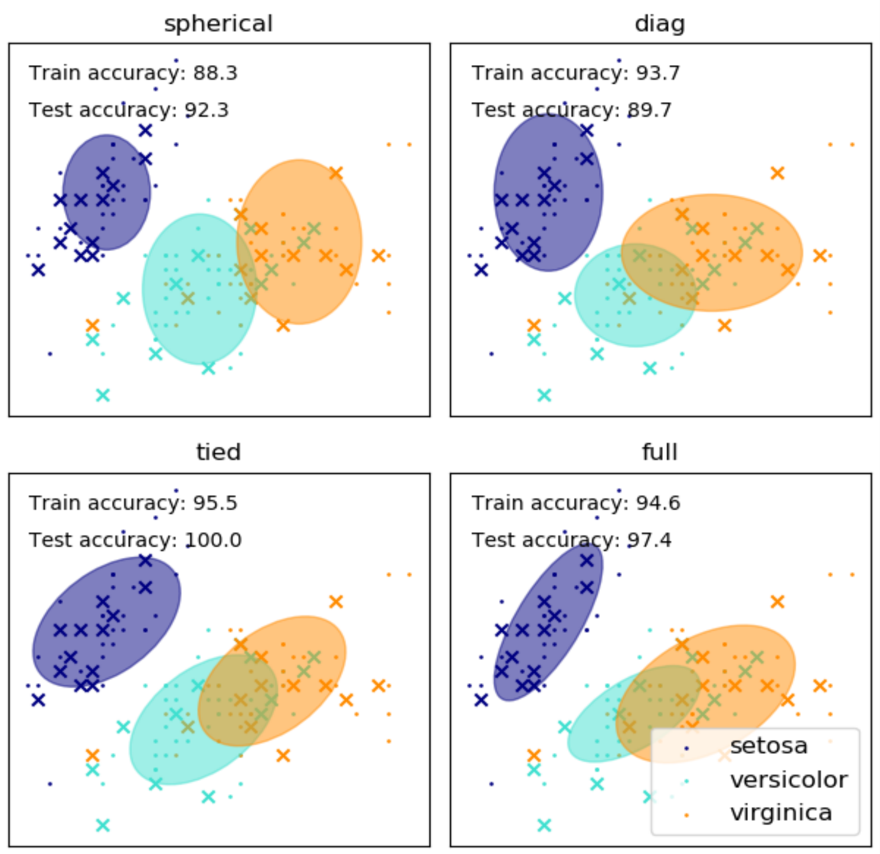


Photo Courtesy of: http://scikit-learn.org/stable/modules/mixture.html

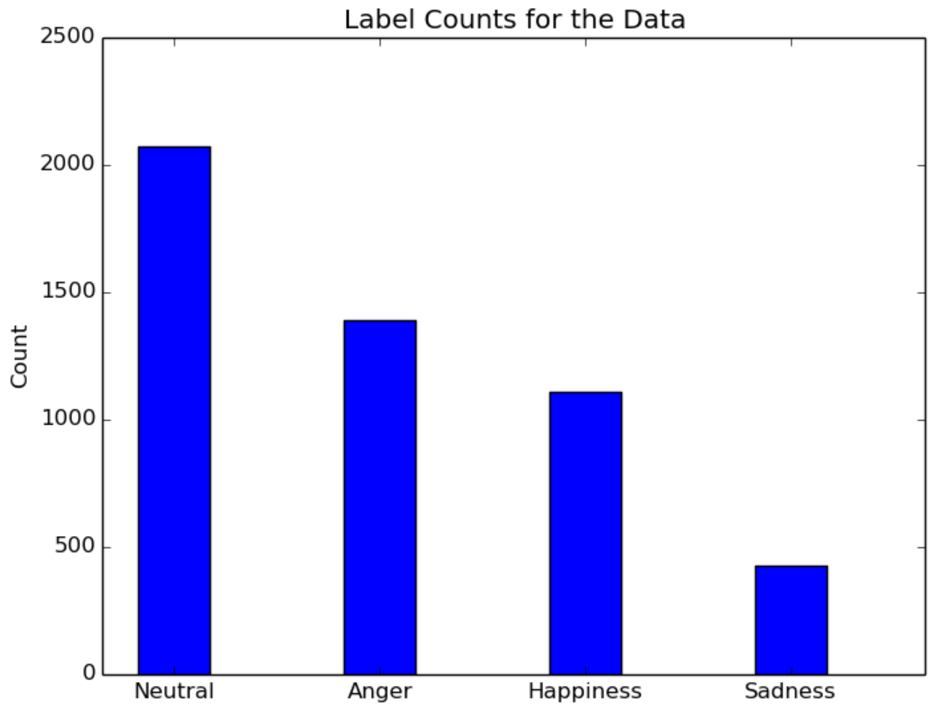
Classification (Continued)

- **Another approach:**
  - Combine similar classes into a single, larger class
  - 9 classes reduced 4 classes in total

- **Support Vector Machines**
  - Utilized Scikit-learn Python library

- **Classification accuracy with test set:**
  - 44.5% with RBF kernel

Problems



Label Counts for the Data
Count
0
500
1000
1500
2000
2500
Neutral
Anger
Happiness
Sadness

Improvements

- **How can we make emotional classification better?**
  - Better distribution of data
  - More data
  - Better models
  - The combination of visual and audio analysis
    - Visual cues such as body language or facial expression can be very indicative of emotion
  - NLP? Do the words spoken by a person help determine emotional context? Maybe.

The End

Questions?