Hannah Portes
COMS 6998 Advanced Topics in Spoken Language Processing
Homework 3: Error analysis

Best Performing
My best performing model was from leaving out the speaker gg. This produced an accuracy of 0.281 and an F1 score of 0.265.

```
SVM accuracy: 0.28095238095238095
SVM F1 (weighted): 0.2654592056789423

              precision    recall  f1-score   support

     anxiety       0.50      0.50      0.50        30
     boredom       0.47      0.53      0.50        30
  cold-anger       0.16      0.26      0.19        27
    contempt       0.21      0.38      0.27        26
     despair       0.18      0.29      0.22        28
     disgust       0.30      0.06      0.10        51
     elation       0.29      0.46      0.36        28
       happy       0.16      0.20      0.18        30
   hot-anger       0.64      0.41      0.50        22
    interest       0.17      0.23      0.20        30
     neutral       0.33      0.11      0.17         9
       panic       0.48      0.48      0.48        27
       pride       0.19      0.16      0.17        25
     sadness       0.00      0.00      0.00        33
       shame       0.33      0.25      0.29        24

    accuracy                           0.28       420
   macro avg       0.29      0.29      0.28       420
weighted avg       0.29      0.28      0.27       420
```

Easiest to Predict Classes
Anxiety, boredom, hot-anger, and panic were the easiest classes to predict. The model achieved an F1 score of over 0.48 - 0.5 for these four classes. The next highest score is 0.36, so there is a significant space between these four top classes and the rest in the dataset. This experiment trained on 1904 samples and tested on 420 samples. The distribution of these classes in the test data is shown in the classification report.

Why were these easy
Anxiety, *boredom*, *hot-anger*, and *panic* each have 30, 30, 22, and 27 samples in the test data respectively. In the training data, *anxiety* had 140 samples, *boredom* had 124 samples, *hot-anger* had 117 samples, and *panic* had 114 samples. This train-test split is the closest to the 80:20 ratio that is recommended typically (the exact percentage is 18.07% of the data is test data). I am inclined to believe that this is partially the reason for this test experiment being the highest performing out of all seven speakers. Additionally, the a*nxiety*, *boredom*, *hot-anger*, and *panic*,

recordings have distinct distributions of normalized intensity and pitch which I believe helps with the accuracy of the classifier.

Which were most difficult

The classes most difficult to predict were *sadness*, with an F1 score of 0.0, *disgust*, with an F1 score of 0.10, *neutral*, and *pride*, both with scores of 0.17. In the training data, *sadness* had 118 samples, *neutral* had 70 samples, *disgust* had 121 samples, and *pride* had 125 samples. The low proportion of training data for neutral specifically does not appear to have impacted the training, as *neutral* performs just as poorly as *sadness*, which had almost the same number of training samples as the higher performing *pride*.

Why were these difficult

Contrasting the easier to predict classes listed above, the normalized values for *sadness, disgust, neutral,* and *pride* are not distinct and could be encompassed by other emotion classes by the classifier. The number of samples as well as the non-distinct extracted features I believe negatively impact the models performance for these classes.

Ideas to improve the classifier

To improve this classifier I would be interested in collecting more data samples from more speakers, as acquiring more data can alone help improve model performance. I would also be interested in exploring more cross validation as opposed to Leave One Speaker Out CV to see if this method of cross-validation had a significant effect on the models performance. Lastly I would explore different normalization techniques, perhaps my method of normalizing each feature by speaker instead of lumping classes together (such as pcm_RMSenergy_sma_max and pcm_RMSenergy_sma_kurtosis) had a negative effect on the accuracy and f1-score of the SVM model I used in this experiment.