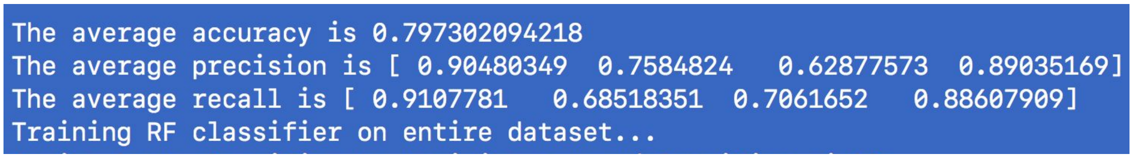
A3 Speaker Identification Writeup

Shivangi Singh, Brittany Pine, Ruifeng Wang, Eric Atwood

1. Describe the best classifier model you used and report its average precision, recall, and accuracy over the 10 folds. Also describe which features you implemented on top of the formants and pitch data (mean, variance, etc.).

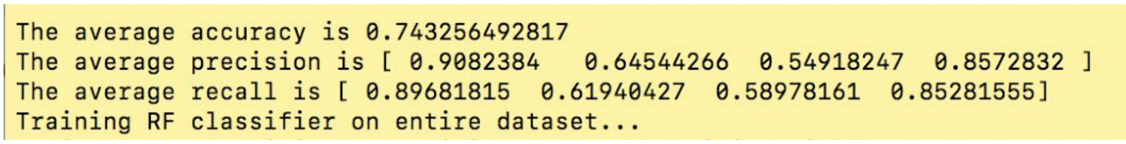
We have finalized our features to delta coefficients, formant distribution (6 bins), delta of delta coefficients, pitch distribution (18 bins), and mean bandwidth. The random forest classifier with n=1000 works best, but may be overfitting. Therefore, we have decided to use the random forest classifier with n=500. The order of the data is none, Brittany, Shivangi, and Ruifeng.

Random forest Classifier (n=1000)



```
The average accuracy is 0.797302094218
The average precision is [ 0.90480349  0.7584824   0.62877573  0.89035169]
The average recall is [ 0.9107781   0.68518351  0.7061652   0.88607909]
Training RF classifier on entire dataset...
```

Random Forest Classifier (n=500)



```
The average accuracy is 0.743256492817
The average precision is [ 0.9082384   0.64544266  0.54918247  0.8572832 ]
The average recall is [ 0.89681815  0.61940427  0.58978161  0.85281555]
Training RF classifier on entire dataset...
```

2. How did you collect your data? Describe the acoustic environments and any other factors that may have an effect on your classifier. How do you account for variation?

We collected audio data at Worcester Dining Commons and the Science and Engineering Library. This allowed us to have locations with varying levels of background noise. All speakers held the phone about 6 inches from their face. When there was no speaker, the phone was placed on the table. All audio samples were 3 minutes long. We tried to account for variation by getting samples in different places, using multiple speakers, and having multiple trials for speakers (in different locations or situations).

Audio Data	Sample 1	Sample 2
No Speaker	Worcester (most background noise)	Science Library (quiet)
Brittany	Worcester (talking in background but less than no speaker)	Science Library (quiet)
Shivangi	Worcester (talking in background but less than no speaker, a metal tray was dropped halfway and produced a loud noise)	Science Library (quiet)
Ruifeng	Science Library (quiet)	Science Library (interviewed by Shivangi and Brittany)

3. Describe another way (besides the one mentioned in the Background section) that speaker identification can be used for health analytics.

Speech analytics can be used to identify health problems (such as Parkinson's Disease or mental health disorders) and mood. If a family has an Amazon Echo in their house, it could possibly identify speakers and whether or not a person's speech has changed over time. For example if somebody has Multiple Personality Disorder, the disorder may be identified based on different voices by the same person.

More of a reach (because everybody's voice would need to be kept in a database), if somebody is asking for help on the phone, the person could possibly be identified.

4. Josh claims that one's vocabulary might be informative enough to distinguish speakers; for instance, he might say "wicked" once in awhile, but his mom still thinks of witches when she hears the word. Assuming that speech recognition tools are as good as we are at hearing, what problem might you still face when trying to implement this feature for speaker recognition? (There is an extra credit part where you can try it!).

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

We can possibly recognize the speaker from the syntax of their speech, as people have their own style of speaking. We could use the N-gram models to find out the words with the highest probability of the nearest words and see if it matches with our speaker. Some drawbacks could be that:

- Households may have similar speech patterns leaving the people hard to distinguish based on syntax alone. (It would probably be easier to identify who it was using voice as well.)
- We might be limited in terms of identifying people when they are reading other people's speech this method can only be used for day to day interactions.

Appendix - Progression Towards Final Classifier

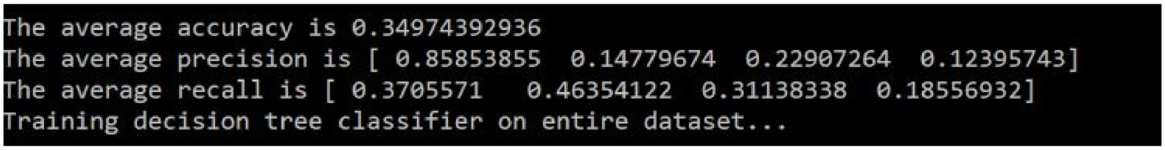
Discuss the changes you made to improve accuracy. How well did things work? What worked best? Give concrete accuracy / precision / recall numbers so we have a sense of how much your classification improved from the baseline.

The following is our progression towards our best classifier model.

<u>1. This is the progression of feature addition.</u>

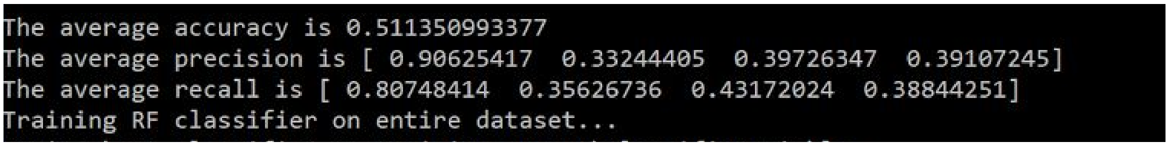
Delta coefficients and formant (6 bins) features (baseline):

Decision Tree (max_depth=3)



```
The average accuracy is 0.34974392936
The average precision is [ 0.85853855  0.14779674  0.22907264  0.12395743]
The average recall is [ 0.3705571   0.46354122  0.31138338  0.18556932]
Training decision tree classifier on entire dataset...
```

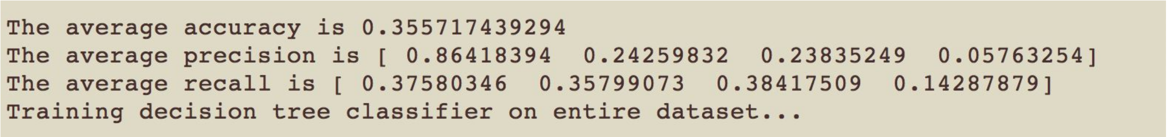
Random Forest (n=100)



```
The average accuracy is 0.511350993377
The average precision is [ 0.90625417  0.33244405  0.39726347  0.39107245]
The average recall is [ 0.80748414  0.35626736  0.43172024  0.38844251]
Training RF classifier on entire dataset...
```

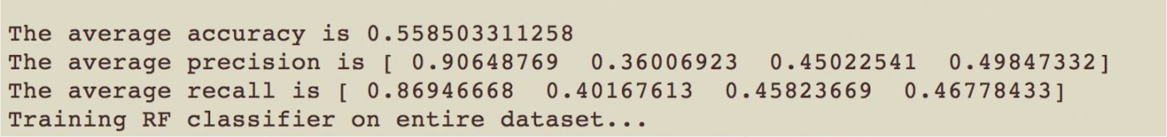
Delta coefficients, formant (6 bins), and delta of deltas as features:

Decision Tree (max_depth=3)



```
The average accuracy is 0.355717439294
The average precision is [ 0.86418394  0.24259832  0.23835249  0.05763254]
The average recall is [ 0.37580346  0.35799073  0.38417509  0.14287879]
Training decision tree classifier on entire dataset...
```

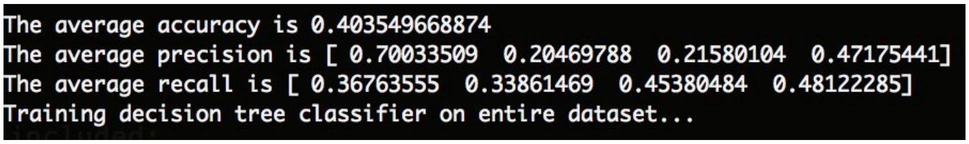
Random Forest (n=100)



```
The average accuracy is 0.558503311258
The average precision is [ 0.90648769  0.36006923  0.45022541  0.49847332]
The average recall is [ 0.86946668  0.40167613  0.45823669  0.46778433]
Training RF classifier on entire dataset...
```

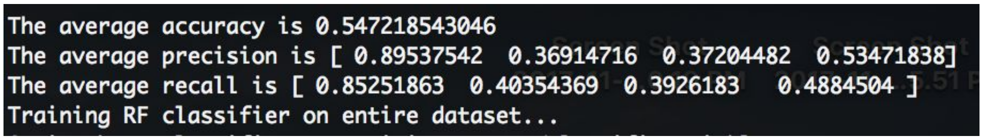
Using delta delta feature but changed histogram bins from 6 to 5 for formants:

Decision Tree (max_depth=3)



```
The average accuracy is 0.403549668874
The average precision is [ 0.70033509  0.20469788  0.21580104  0.47175441]
The average recall is [ 0.36763555  0.33861469  0.45380484  0.48122285]
Training decision tree classifier on entire dataset...
```

Random Forest (n=100)

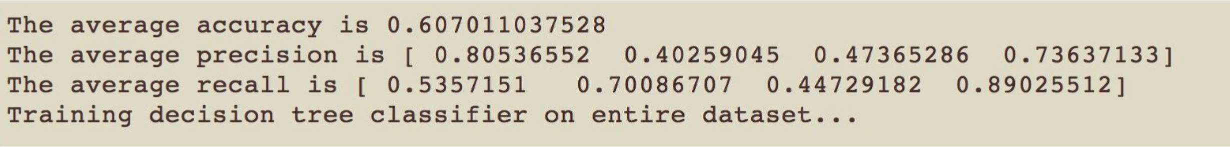


```
The average accuracy is 0.547218543046
The average precision is [ 0.89537542  0.36914716  0.37204482  0.53471838]
The average recall is [ 0.85251863  0.40354369  0.3926183   0.4884504 ]
Training RF classifier on entire dataset...
```

As seen here, 6 bins gave better results than 5.

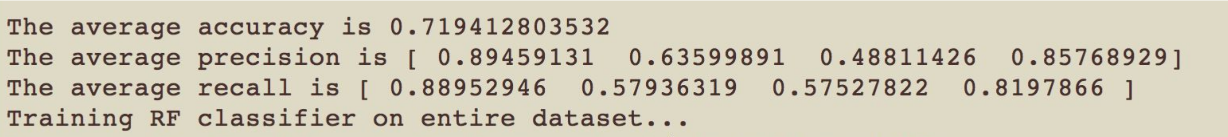
We chose 18 bins in the histogram for pitch because the plotted histograms looked best: data was spread out enough so that it was in multiple bins but not so much that there were many bins with 1 or 0 datum.

Delta coefficients, formant (6 bins), delta of deltas, and pitch (18 bins) as features:
Decision Tree (max_depth=3)



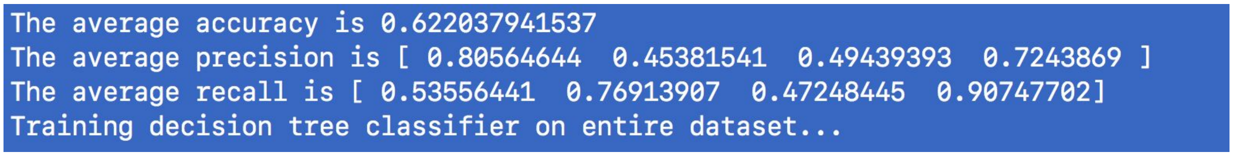
```
The average accuracy is 0.607011037528
The average precision is [ 0.80536552  0.40259045  0.47365286  0.73637133]
The average recall is [ 0.5357151   0.70086707  0.44729182  0.89025512]
Training decision tree classifier on entire dataset...
```

Random Forest (n=100)



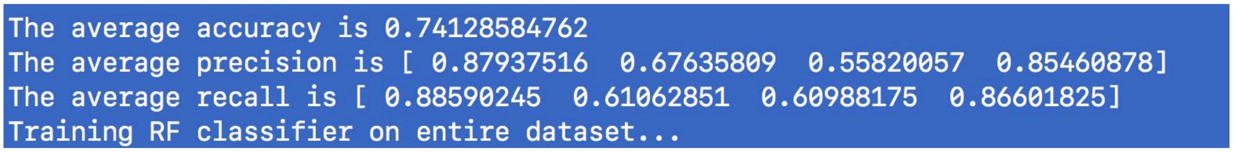
```
The average accuracy is 0.719412803532
The average precision is [ 0.89459131  0.63599891  0.48811426  0.85768929]
The average recall is [ 0.88952946  0.57936319  0.57527822  0.8197866 ]
Training RF classifier on entire dataset...
```

Delta coefficients, formant (6 bins), delta of deltas, pitch (18 bins) and mean bandwidth as features:
Decision Tree (max_depth=3)



```
The average accuracy is 0.622037941537
The average precision is [ 0.80564644  0.45381541  0.49439393  0.7243869 ]
The average recall is [ 0.53556441  0.76913907  0.47248445  0.90747702]
Training decision tree classifier on entire dataset...
```

Random Forest (n=100)

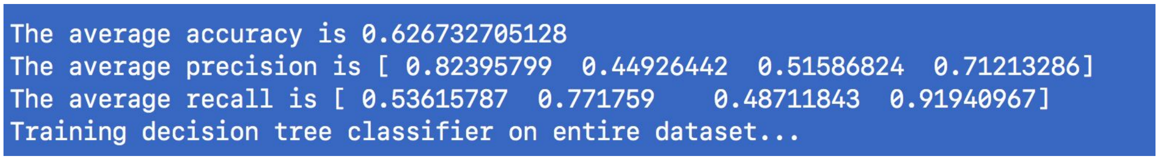


```
The average accuracy is 0.74128584762
The average precision is [ 0.87937516  0.67635809  0.55820057  0.85460878]
The average recall is [ 0.88590245  0.61062851  0.60988175  0.86601825]
Training RF classifier on entire dataset...
```

We have finalized our features to delta coefficients, formant (6 bins), delta of deltas, pitch (18 bins) and mean bandwidth.

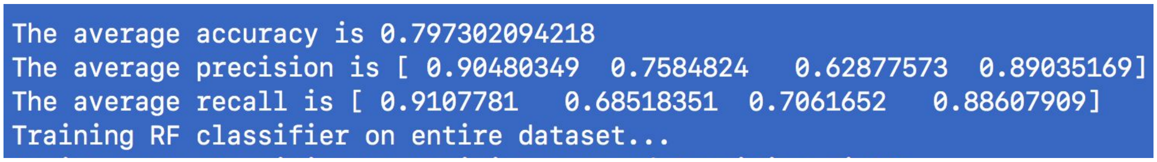
2. The following is choosing between the best classifier parameters.

Decision Tree: (max_depth=3)



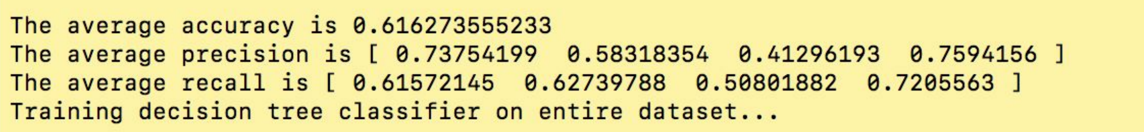
```
The average accuracy is 0.626732705128
The average precision is [ 0.82395799  0.44926442  0.51586824  0.71213286]
The average recall is [ 0.53615787  0.771759    0.48711843  0.91940967]
Training decision tree classifier on entire dataset...
```

Random Tree Classifier (n=1000)



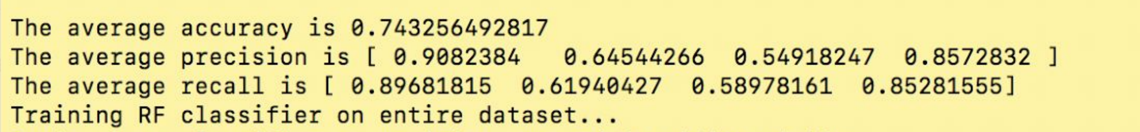
```
The average accuracy is 0.797302094218
The average precision is [ 0.90480349  0.7584824   0.62877573  0.89035169]
The average recall is [ 0.9107781   0.68518351  0.7061652   0.88607909]
Training RF classifier on entire dataset...
```

Decision Tree (max_depth=8)



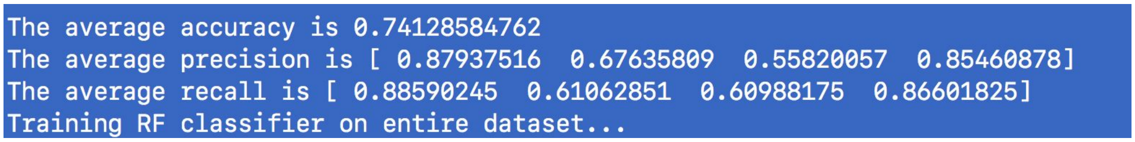
```
The average accuracy is 0.616273555233
The average precision is [ 0.73754199  0.58318354  0.41296193  0.7594156 ]
The average recall is [ 0.61572145  0.62739788  0.50801882  0.7205563 ]
Training decision tree classifier on entire dataset...
```

Random Forest (n=500)



```
The average accuracy is 0.743256492817
The average precision is [ 0.9082384   0.64544266  0.54918247  0.8572832 ]
The average recall is [ 0.89681815  0.61940427  0.58978161  0.85281555]
Training RF classifier on entire dataset...
```

For reference, this is repeated from above.
Random Forest (n=100)



```
The average accuracy is 0.74128584762
The average precision is [ 0.87937516  0.67635809  0.55820057  0.85460878]
The average recall is [ 0.88590245  0.61062851  0.60988175  0.86601825]
Training RF classifier on entire dataset...
```

The random forest classifier with n=1000 works best, but we think this may be overfitting the data. Therefore, we have decided to use the random forest classifier with n=500.