Human Speech Recognition

Gregory Lull

Motivation

Use cases

- Interaction with smart devices
- Forensic analysis
- Emergency channels

Dataset and methodology

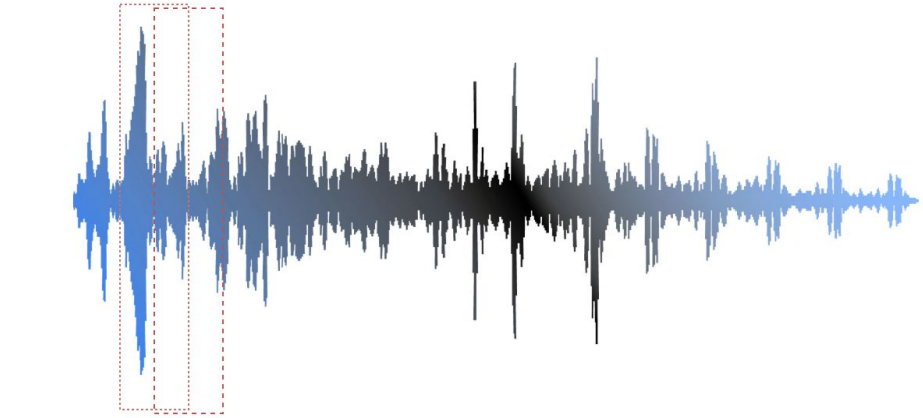
Data:

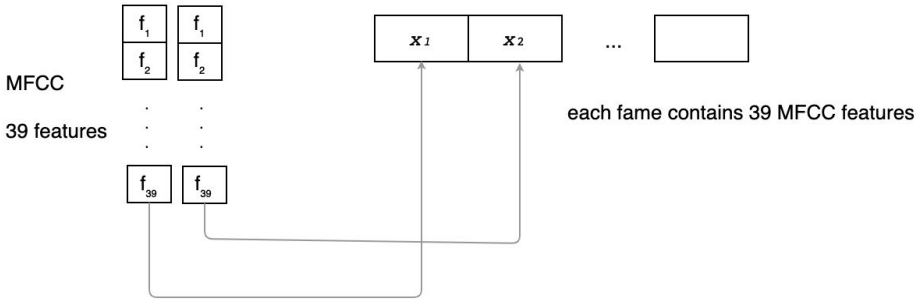
- CHiME Speech Separation and Recognition Challenge
- N = ~2000 validated, 4 second clips
- Somewhat balanced: 65% human voices, 35% ambient noise

Methods:

- Extract features: Mel Frequency Cepstral Coefficient (MFCC)
- PCA to reduce features from thousands to dozens
- Modeling with cross-validation and parameter tuning
    - Logistic Regression, KNN, Random Forest, SVM

Feature extraction: Mel Frequency Cepstral Coefficient (MFCC)





Results - KNN using F1 scores

F1 Score
Naive model: 80.6%
KNN model: 87.8%
Improvement: ~ 9%

		Predicted	
		Has human	No human
Actual	Has human (83.7% correct)	241	22
	No human	45	82

Conclusion

- Unfortunately using PCA makes the model less interpretable, however it still demonstrates that ML methods can make a difference

Future Work

- Larger dataset:
  - mine was 4gb with 2000 samples, there are free datasets online with millions of audio samples
- Finding a more diverse and balanced set, such as nature sounds, music, shouting, etc.
- Apply this to a long audio clip and validate where speech occurs.

Appendix - PCA and KNN best params, other params

KNN n_neighbors: 25
PCA n_components: 0.2

Tuning params for other models:
Logreg

```
'logreg__C': np.arange(1, 1000, 100),
'pca__n_components': pca_component_range
```

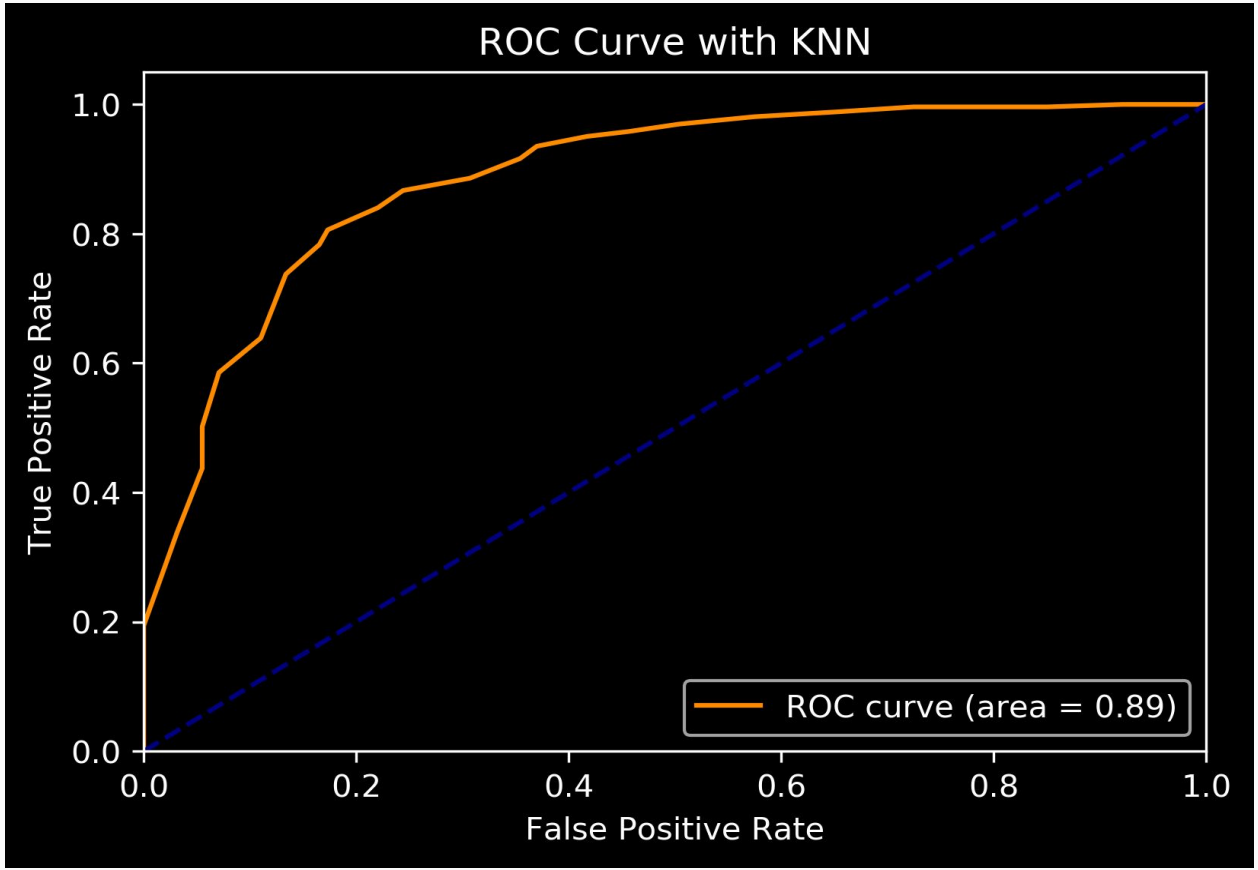
SVM

```
'svm__kernel': ['poly', 'rbf'],
'svm__C': np.arange(1, 3, 0.25),
'svm__degree': [3, 4],
'pca__n_components': pca_component_range
```

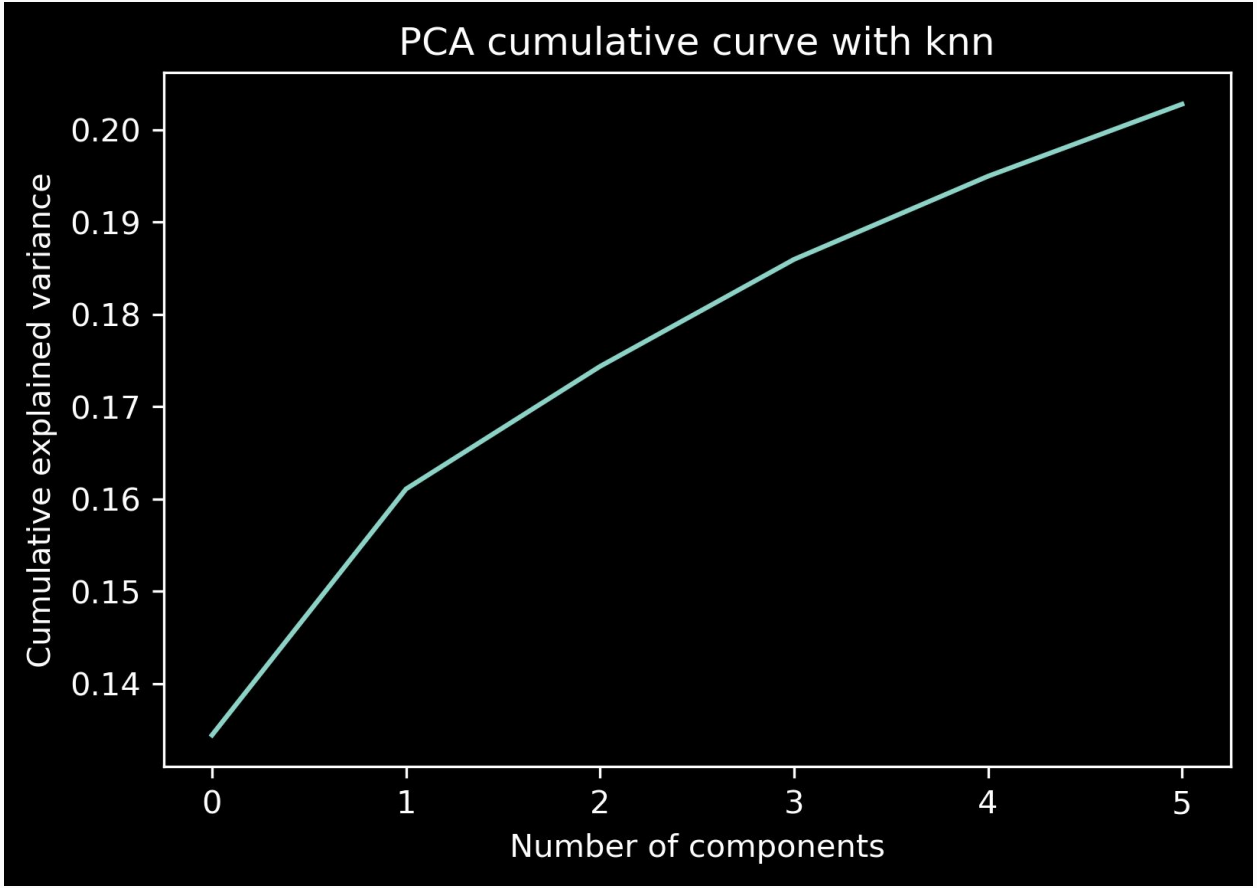
RandomForest

```
'pca__n_components': pca_component_range,
'randomforest__n_estimators': np.arange(10, 150, 15)
```

Appendix - KNN ROC Curve



Appendix - KNN PCA cumulative curve



Appendix - PCA cumulative curve for features only (before model fitting)