ABLoc

Audio-Based Localization

James Sun and Reid Westwood

Motivation

As people go about their daily routine, often they begin to recognize their location just by the sounds they hear. We aim to test if a machine can do the same by investigating the distinctness of soundscapes between locations on campus. If actually possible, this ability could augment traditional localization methods with qualitative details.

Much of the literature on audio-based learning has focused on speech recognition. However, interest in using audio for broader applications is increasing. Previous related work has included audio-augmented scene recognition for robotics [1] and sound type discrimination [2]

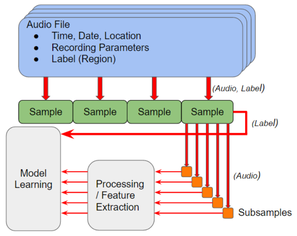
[1] S. Chu, S. Narayanan, C. c. J. Kuo, and M. J. Mataric, “Where am i? scene recognition for mobile robots using audio features,” in 2006 IEEE International Conference on Multimedia and Expo,
July 2006, pp. 885–888.
[2] L. Chen, S. Gunduz, and M. T. Ozsu, “Mixed type audio classification with support vector machine,” in 2006 IEEE International Conference on Multimedia and Expo, July 2006, pp. 781–784.

Methodology

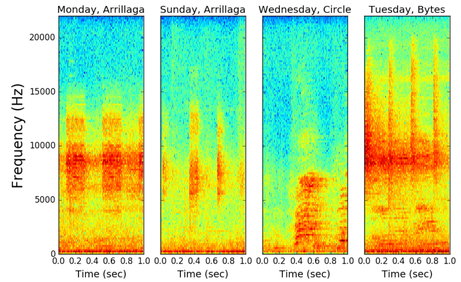
Raw Data Collection

7 Regions
- *Outdoor*: Rains; “Circle of Death”; Huang; Oval
- *Indoor:* Tressider; Bytes; Arrillaga Gym

1 Minute recordings at a time



Methodology

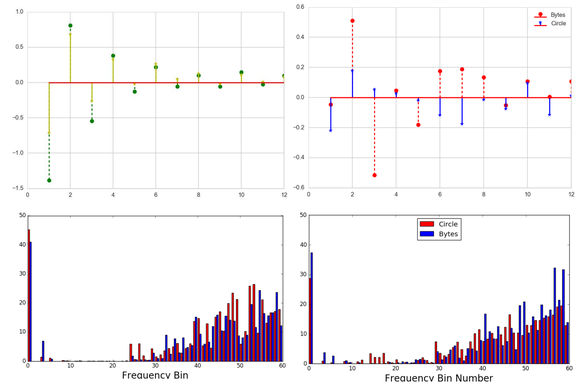


Sample Spectrograms

Audio Features

MFCC
- Measure of power in short term spectrum of a signal.
- Mimics human hearing

Spectrogram Peak Detection (SPD)
- Counts local maxima in frequency
- Designed to detect consistent energy bands over time



Arrillaga, 2 Days *Bytes, Circle*

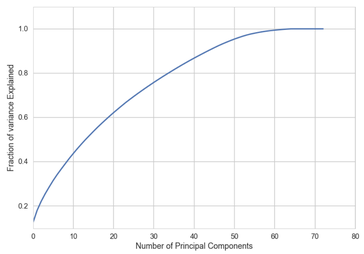
Results

Summary

As seen in the classifier comparison, our ensemble method, which used a Gaussian kernel SVM as a primary classifier and the linear logistic as a secondary classifier, produced the most promising results. When testing data, we settled on using 10 subsamples of 1 second audio. While the test error was reduced by using more data, we found that this duration offered a good tradeoff with application – a user may not spend much more than 10 seconds in a given spot (or want to hold their phone out for longer).

Initial results are very encouraging. However, gathering representative data remains a large challenge given the hugely temporal nature of our dataset. Classifier interpretation also poses difficulties given our feature space’s dimensionality and nature. Nevertheless, we are optimistic that the increasing availability of data and ubiquity of technologically advanced personal devices can greatly expand the scope of this project and allow for its integration in general localization systems.

We chose a set of 73 features for each audio sample (13 MFCCs and 60 SPDs). Using PCA, we saw that a large subset of these features were needed to explain the variation in our dataset. Of the 73 principal components, the first 50 accounted for 95% of the data variance.



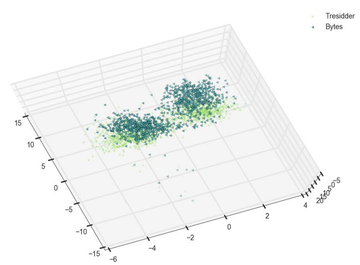
Confusion Matrices

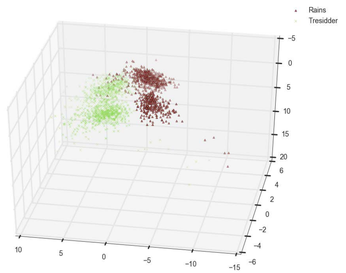
	Rains	Circle	Tressider	Huang	Bytes	Oval	Arrillaga
Rains	0.9661	0.0169	0.	0.0169	0.	0.	0.
Circle	0.0638	0.8298	0.	0.0426	0.0213	0.0426	0.
Tressider	0.	0.	1.	0.	0.	0.	0.
Huang	0.	0.	0.	1.	0.	0.	0.
Bytes	0.	0.	0.	0.	0.9778	0.	0.0222
Oval	0.0889	0.1778	0.	0.0222	0.	0.7111	0.
Arrillaga	0.	0.	0.0233	0.	0.	0.0233	0.9535

X-Validation Error

	Rains	Circle	Tressider	Huang	Bytes	Oval	Arrillaga
Rains	0.85	0.1167	0.	0.	0.	0.0333	0.
Circle	0.0333	0.7833	0.1	0.	0.05	0.0333	0.
Tressider	0.	0.082	0.7705	0.	0.0656	0.	0.082
Huang	0.	0.1167	0.0167	0.85	0.	0.0167	0.
Bytes	0.	0.0667	0.05	0.	0.8833	0.	0.
Oval	0.0333	0.35	0.0167	0.0333	0.	0.4333	0.1333
Arrillaga	0.	0.0909	0.0152	0.0303	0.0455	0.0152	0.803

Generalization Error





We used the first 3 principal components as a basis to visualize our data. Using just these 3 principal components, we were able to visually see clear separations in some pairs of regions, such as Rains and Tressider (Right). However, other region pairs did not have quite so nice a separation in this low dimensional projection, such as Tressider and Bytes (Left).

Classifier Comparison

Unsurprisingly, we found that data gathered within a single day at a given location is highly correlated, compared to between different days. Because of this, we measure our test classification methods two ways.

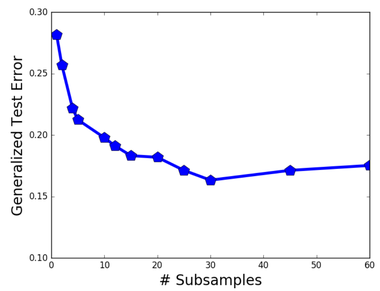
First, standard cross correlation was used, where each sample was treated independently.

Second, which we refer to as ‘generalization error’, we hold out all data gathered on a single day from training. This day’s data is then our test data.

	X-Validation Error	Generalization Error
Gaussian Kernel SVM	13.65%	21.72%
Linear SVM	27.84%	32.74%
Linear Logistic	15.45%	21.22%
Random Forest	14.09%	28.26%
RBF+Logistic Ensemble	13.89%	19.68%

Voting

Because we use 1-second processing to generate features, just 10 seconds of data give us 10 test samples known to have a single label. This curve shows the decrease in test error as we increase the duration of a test clip.



Our use of ensemble takes advantage of need for voting. The process is as follows:

1. Use our primary classifier to predict each test subsample
2. If the gap between 1st and 2nd is small, run the test sample through the secondary classifier
3. Use a weighted average of the two classification results to determine prediction.