Campus Location Recognition using Audio Signals

James Sun,Reid Westwood
SUNetID:jsun2015,rwestwoo
Email: jsun2015@stanford.edu, rwestwoo@stanford.edu

I. Introduction

Recognizing one's location by sound is a coarse skill that many people seem to develop out of routine. We may be able to recognize a favorite café by the genre of music playing and the baristas' voices. We may be able to recognize the inside of our car by the noises coming out of the engine and chassis. We might come to associate the sounds coming through our rooms' windows with home. However, are these sounds by themselves truly sufficient to identify the locations that we frequent? This project attempts to answer that question by developing a Machine Learning system that recognizes geographical location purely based on audio signal inputs. To emulate a typical Stanford student, the system is trained on sounds at locations along a path that a student might take as he or she goes about a typical school day. In the process of developing this system, we investigated audio features in both the spectral and time domain as well as multiple supervised learning algorithms.

II. Related Work

A previous CS229 course project identified landmarks based on visual features [1]. [2] gives a classifier that can distinguish between multiple types of audio such as speech and nature. [3] investigates the use of audio features to perform robotic scene recognition. [4] integrated Mel-frequency cepstral coefficients (MFCCs) with Matching Pursuit (MP) signal representation coefficients to recognize environmental sound. [5] uses Support Vector Machines (SVMs) with audio features to classify different types of audio.

III. Scope

As stated in Section I, we have limited the number of areas that the system will recognize. Furthermore, we have limited the geographical resolution of labels to named locations encompassing areas such as Rains Graduate Housing. Both of these limitations are in line with how a typical person may use audio cues to identify his or her location. As such, these geographical restrictions in scope are unlikely to be relaxed.

We have also initially limited our scope temporally to data gathered on weekdays between the periods of 9AM to 5PM during the Spring Academic Quarter. Initial results are promising, and we may relax some of these restrictions.

IV. System Design

A. Hardware and Software

The system hardware consists of an Android phone and a PC. The Android phone runs the Android 6.0 Operating system and uses the `HI-Q MP3 REC (FREE)` application to record audio. The PC uses Python with the following open-source libraries:

- Scipy
- Numpy
- statsmodels
- scikits.talkbox
- sklearn

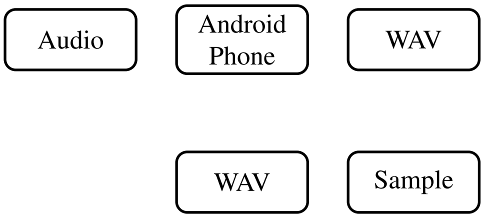
The system also makes use of a few custom libraries developed specifically for this project.

B. Signal Flow

The following details the flow of a signal when making a prediction

1) Audio signal is recorded by the Android phone
2) Android phone encodes the signal as a Wav file
3) The Wav file enters the Python pipeline as a `Sample` instance
4) A trained `Classifier` instance receives the `Sample`
    a) The `Sample` is broken down a number of subsamples based on a predetermined audio length for each subsample
    b) A prediction is made on each subsample
    c) The most frequent subsample prediction is output as the overall prediction.

A graphical illustration of this is below:



C. Locations

The system is trained to recognize the following 7 locations:

0. Arrillaga Gym
1. Bytes Café
2. Circle of Death
   Intersection of Escondido and Lasuen
3. Huang Lawn
4. The Oval
5. Rains Graduate Housing
6. Tressider Memorial Union

These locations represent the route a typical graduate engineering student living at Rains might take on a typical day. Locations 0,1, and 6 are indoors whereas Locations 2,3,4, and 5 are outdoors.

V. Data Collection

A. Format

Data is collected using the `HI-Q MP3 REC (FREE)` application as noted in Section IV-A. This application is freely available on the Google Play Store. Monophonic Audio is recorded without preprocessing and postprocessing at a sample rate of 44.1 kHz.

B. Training Data

Training data is gathered during weekdays in the morning in order

VI. Methods

The goal is to have the system recognize natural geographic aggregates rather than recognize individual fine-grain coordinates. For example, the system will label a data set as belonging to ”The Quad” or ”Bytes Cafe”, similar to how a person would naturally describe an environment. I expect to use supervised learning algorithms to separate data points based on audio and visual features.

A. Audio Features

As audio is potentially the more interesting data type, I have come up with a few basic features to evaluate. These include the following:

- Frequency Spectrum Bandwidth
- Frequency Spectrum Variance
- Frequency Spectrum mean
- Intensity variance
- Mean Intensity

References

[1] A. Crudge, W. Thomas, and t. . Kaiyuan Zhu.

[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.

[3] 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.

[4] S. Chu, S. Narayanan, and C. C. J. Kuo, “Environmental sound recognition with time and frequency audio features,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1142–1158, Aug 2009.

[5] G. Guo and S. Z. Li, “Content-based audio classification and retrieval by support vector machines,” *Neural Networks, IEEE Transactions on*, vol. 14, no. 1, pp. 209–215, 2003.