Bird Songs Analysis

Zhengyuan Wen, Ming Pei, Haoyue Shi

Outline

- Introduction
- Data Preprocessing
- What we did on CHTC
- Our classification model
- Future Prospects and drawbacks

Introduction

Our goal: Identify the bird specie given its song

Our final data set: 12GB audio files of 13 different bird species

Our Result: A model with an accuracy 49.14% on test data sets, which is larger than a random guess

(7.7%) among 13 bird species

Data Preprocessing

In this part, we tried to **change audio files to a “dataframe”**, which can be directly used by the classification model

Data Preprocessing

Step 1

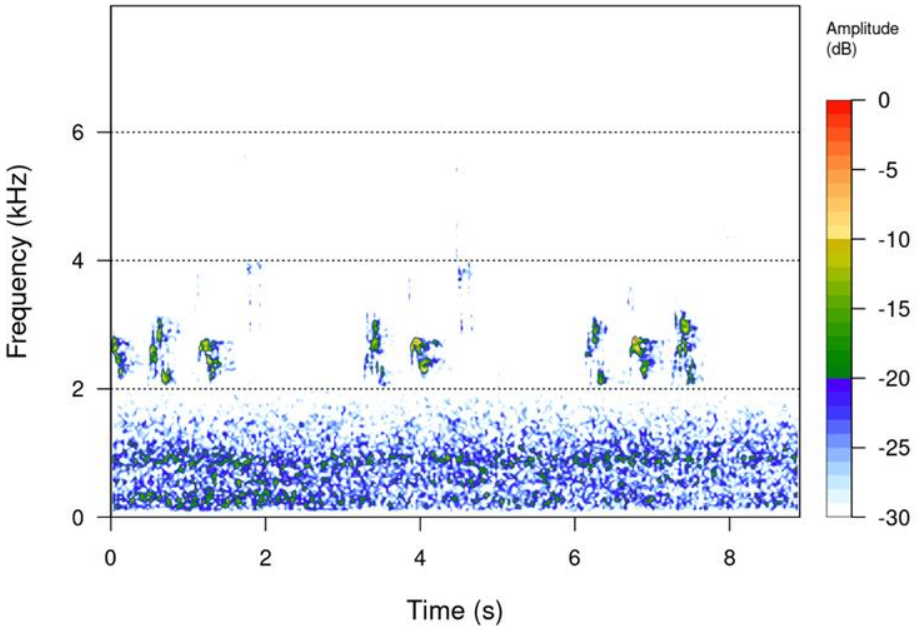
Goal: Remove the background noises. The high frequency is usually the sound of birds, and the low frequency is usually some environmental noise.

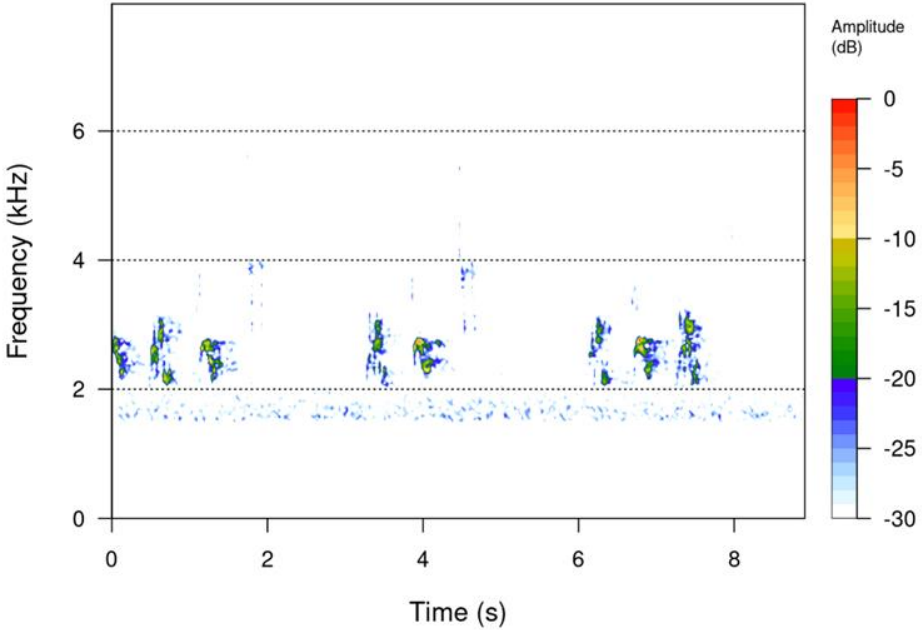
Method: FIR(Finite Impulse Response filter). We manually set a frequency baseline 1500Hz by listening to different birdsongs of 13 bird species. And remove the amplitudes whose frequency are smaller than 1500Hz.

Data Preprocessing

Step 1

Results:





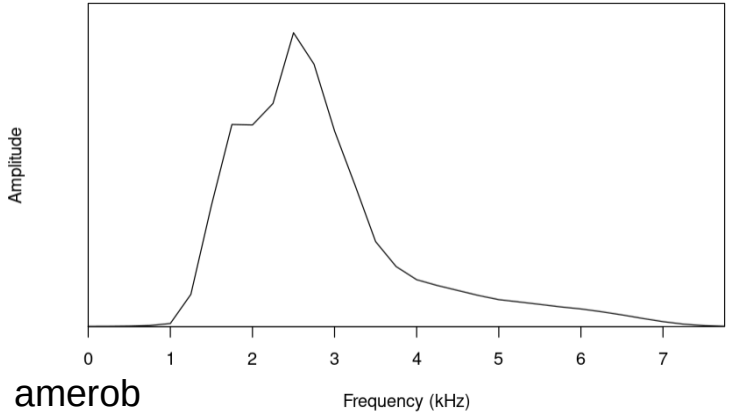
Data Preprocessing

Step 2

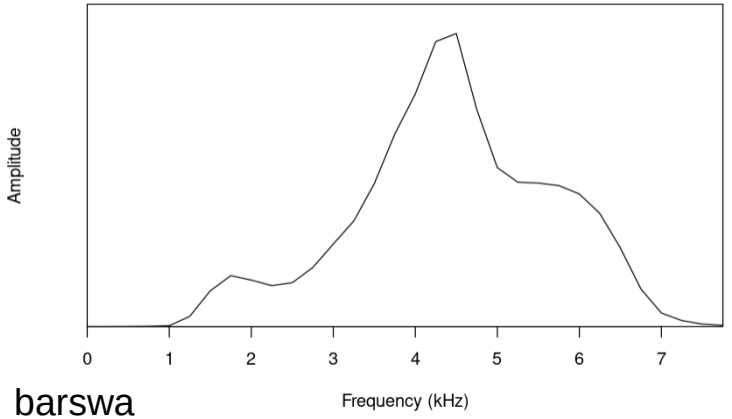
Goal: Extract features from audio files

Method: Extract the mean relative amplitude of the frequency distribution **using STFT(Short-time Fourier Transformation)**

Results: We got 32 features extracted from an audio file



amerob



barswa

What we did on CHTC

We used CHTC to do the data preprocessing jobs

- The total number of samples is 6727. We used sample() function in R to obtain **30 groups of samples** and **deployed on 30 CHTC machines.**
- Each group contains roughly 225 files and each of their sizes is about 350MB. Since FIR requires large amount of memories, each procedure we requested **1CPU (Used 1), 2GB of Disk Space (Used roughly 1.3GB) and 8GB of memories (Used roughly 6GB).**
- **A single job can be completed in 14 minutes.** The total 30 parallel jobs were completed in 1 hour, saving us a lot of time.

Model introduction

- For training part, we used **Single Hidden-Layer Neural Network with skip layer connection** for prediction.
- We tried several parameter options for size of the hidden layer, and ended up **choosing 8 knots in hidden layer** as our final choice. Larger ones (16 and 12) showed worse results than the 8 because of overfitting problem.
- The total parameters (weights) we need to estimate are 797 ($32\times8+8\times13+32\times13+13+8=797$).

Classification Model

Predict example

	amerob	barswa	bewwre	blujay	comrav	comter
150	0.0004219843	0.2216296302	0.05247639	5.070655e-06	0.001121790	2.765061e-05
421	0.1021005316	0.0004525987	0.03843799	4.095615e-01	0.096949481	1.885261e-01
1921	0.0009750648	0.3226119735	0.03050319	4.574655e-05	0.004380047	4.441303e-04

	daejun	eursta	grycat	houspa	houwre	mallar
150	0.017199828	0.01564742	0.00200396	0.583761271	0.10278093	6.528468e-05
421	0.013098746	0.03005486	0.03764553	0.002770938	0.02585903	5.097268e-02
1921	0.008776813	0.03029090	0.00464660	0.333243792	0.26051183	2.308696e-05

	marwre
150	0.002858783
421	0.003570079
1921	0.003546828

Sample 150's true label is houspa, Sample 421 is blujay and 1921 is houwre.

Model result

The test accuracy is 49.14% similar to **training accuracy (51.40%)**. We say there is no obvious
overfitting problem. (Training data: Group 4-30 / Testing data: Group 1-3)

As for multiple class prediction task, we often use **Cohen's Kappa Coefficients** to measure good or bad of a model.

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = 0.4225$$

where p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement. In this case, $p_e = \text{sum}(\text{ Number of prediction on } i^{th} \text{ bird} \times \text{Number of } i^{th} \text{ bird }) / \text{ Sample\_size }^2$.

$\kappa = 0.4225$ **represents a moderate agreement for categorical data.**

Future prospects and drawbacks

- Hard to distinguish the bird songs whose frequency is lower than 1500 Hz as we didn’t have those birds in training set after selections
- The accuracy of model can be improved by using Convolutional Neural Network and MFCC
- We didn’t deal with the data imbalance issue. (The larges number of samples of certain bird species is 1206 and the samllest is 223)