Autonomous Recognition of Birds from their Calls

Ankita Pasad
Electrical Engineering Department
IIT Bombay

Guide: Prof. Preeti Rao

November 17, 2015

Acknowledgement

I would like to express my sincere gratitude towards my respected guide Professor Preeti Rao for providing us her ideas and support during the course of the project along with the required academic and technical background for the project. I am also thankful to Parthe Pandit, Raunak Bhinge and Pallavi Hujband for their valuable suggestions and encouragement, and to my colleague Riddhish Bhalodia for his cooperation throughout.

1 Abstract

Syllables are elementary bulding blocks of a bird call. Our attempt is to devise an algorithm for automatically recognising the birds from their calls with maximum accuracy. Experimental results of syllable based classification wherein different birds are identified by differnce in various features of their syllable(s) have shown that the proposed method can be efficiently used for bird identification from its recorded call. In this article we test how well bird species can be identified by having different cluster for each bird, followed by testing - to which cluster does the recorded call's syllable fit the best.

2 Introduction

Automatic recognition of birds would prove to be a very helpful tool for further research in ornithology and bird species conservation by acquiring real time and accurate data for recording and monitoring the presence and movements of different bird species. If bird sound analysis is taken to a next level, we can also use it to detect the number of bird species in a particular region by just having a walk covering the entire region with audio recording on. Thus technology for sound-based identification of bird species could be a significant addition to the research methodology in taxonomy and monitoring of migration and population in biology. .

Birds produce sound from their syrinx[1]. It is located in the intersection between main bronchi of the lungs and the trachea. Syrinx in birds can feature multiple simultaneous oscillation modes. Thus the temporal variations in the spectrum of bird sounds in typically orders of magnitude faster than in human sound production. Thus a high temporal resolution in the order of milliseconds is needed in the bird sound analysis. Bird sound is divided into four levels: notes, syllables, phrases and song. Syllables were seen to be the most elementary building block and so we used them as our basis of identification than song paterns. This method would work fine with songs as well as individual syllables can be segmented from a song calls and the a set of syllables can be matched from the test data. Also syllable-based recognization will be invariant with variation of song patterns with region and type of the call of same bird. Also, a recording with the mixture of calls of multiple birds can be segmented into syllables and different birds can be identified.

Relatively little has been done in this field previously. We have gone through research work by Aki Harma who has tried to do sinusoidal modelling of syllables and came up with an algorithm for small set of birds. But, a large class of bird sounds are not pure sinusoids and have a clear harmonic spectrum structure. So, in another paper by the same researcher, the author has tried to differntiate birds on the basis of harmonic class it belongs to, as . Both these methods are the ones which can be used as a single rung of classification, thus leading to hierarchial classification. Considering the number of bird species and the huge database, hierarchial classification will need lots of features and the really accurate ones, otherwise a small error in one level of classification will lead to a bird getting totally misclassified.

In this project we have done a syllable based classification using different clustering techniques. No similarities were observed between calls of birds belonging to same family or order. Various spectral and temporal features were analysed and this was done for multiple sample sound files of each bird; and thus similarities for same bird were extracted and so the differences between different birds. The clustering was first attempted using unsupervised clustering technique of k-means clustering and then supervised clustering algorithm called naive-bayes was used. The features used, experiments performed and their accuracy is discussed in the ensuing sections. Till now we have been successful in working with 45 bird species. The work has not yet reched a conclusive stage and we are in the process of exploring better options and fine-tuning the existing algorithm.

3 Approach

3.1 An Overview

Our working platforms included Audacity (audio-editing software), Praat (phonetic software) and Scilab (for computations). Scilab was replaced by Matlab at the later stage. All the bird sounds were manually segmented using Audacity to extract individual syllables from different recordings. Those extracted sound files were then observed in Praat for their spectogram, pitch and intensity. Here is an example of how a sound file looks like in Praat. The blue line indicates pitch curve and yellow line indicates intensity curve. As can be seen, this particular bird has lots of harmonics in its spectogram and the pitch is quite low at around 400 Hz and intensity curve has 2 peaks. Frequency range for pitch and number of intensity peaks for various recordings were observed using Praat and thus similarities and differences were drawn. After this first step of getting acquainted with sound processing and getting a hang of basic features, we moved on to coding in scilab to get these features along with more.

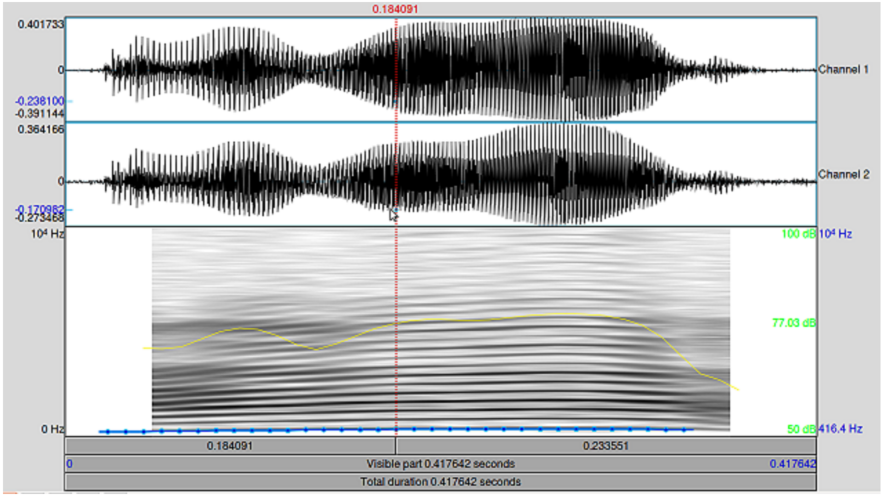


Figure 1: Spectogram of ruddy-shelduck

Following is the broad overview of our approach

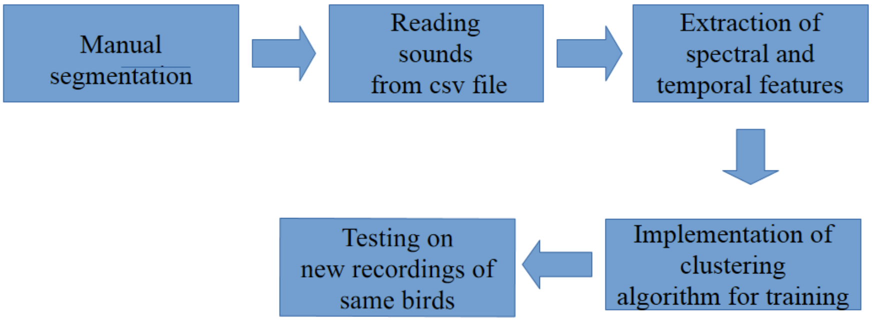

Manual segmentation
Reading sounds from csv file
Extraction of spectral and temporal features
Implementation of clustering algorithm for training
Testing on new recordings of same birds

Figure 2: Block diagram of our approach.

The spectral and temporal features were extracted from the short time fourier transform of the recoring and original sound file respectively.

1. The spectral features included:

    - Spectral Flux
    - Spectral Centroid
    - Spectral Flatness
    - Spectral Rolloff
    - Spectral Crestfactor
    - Spectral Spread

    Here is an example of how these features looked like

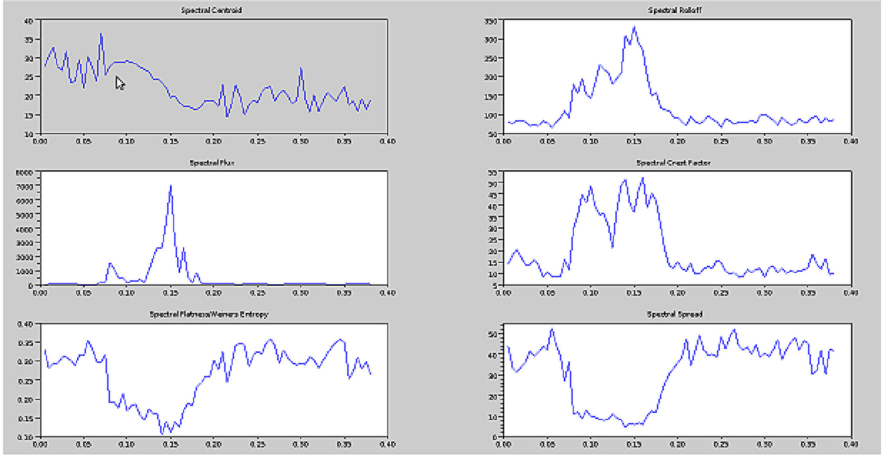


Figure 3: 6 spectral features of Golden eagle

2. The temporal features used are:
    - Syllable duration
    - zero crossing rate

Call duration is observed manually in audacity and ZCR is extracted from scilab.
A csv file is generated with the list of all the extracted features so that the data can be automatically read and processed for further clustering.
Initially unsupervised c-means clustering algorithm was tested on the data, very accurate results were not observed for a big dataset of birds. Also problems like decision on number of clusters and what should be the next rung of classification persisted. We could not go with number of clusters same as number of birds, because in that case all the samples got shuffled up. So, after trying c-means for different iterations we moved on to a supervised clustering algorithm of naive-bayes. Naive Bayes gave quite satisfactory results with accuracy of about 95% after removing a few distinct outliers from the dataset.

3.2 Algorithms

Algorithms and relevant formulae for different parts:

1. Short time fourier transform

   It is a local analysis scheme for a time-frequency representation of a signal. In a nutshell it is segmenting the signals into narrow time intervals followed by taking the fourier transform of each segment. Following are the specifications of the stft calculated by us using the fast fourier transform algorithm.
   Window type: hamming window
   Sampling rate: 48kHz
   Window size: 10 ms
   Window overlap: 5 ms (50%)
   Here is the code snippet for stft function:

   ```
   y=Speech_signal;
   reconsine = y;
   Frame_size=Frame_size/1000;
   Frame_shift=Frame_shift/1000;
   max_value=max(abs(y));
   y=y/max_value;
   //disp(length(y));
   if(length(y) > 48000)
        t=1/Fs:1/Fs:(48000/Fs);
        for p = 1:48000
             y0(p) = y(p);
        end
   else
   ```

```
        t=1/Fs:1/Fs:(length(y)/Fs);
        y0 = y;
end
Frame_length = Frame_size*Fs;
sample_shift = Frame_shift*Fs;

w=window(window_type,Frame_length);jj=1;
dftylast = 0;
for i=1:(floor((length(y))/sample_shift)-
ceil(Frame_length/sample_shift))

    k=1;yy=0;
    for j=(((i-1)*sample_shift)+1):(((i-1)*sample_shift)+Frame_length)
        yy(k)=y(j)*w(jj);
        yyy(k) = y(j);
        jj=jj+1;k=k+1;
    end


    dfty=abs(fft(yy)); // The amplitude information
    dftyp = atan(imag(yy),real(yy)); // The phase information
```

Various features obtained from the stft thus calculated proved to be very useful and are described further.

For 2-7: X(n) is FFT of single frame and 'k' is the half length of FFT.

2. Spectral Centroid

   It is the center point of the spectrum and in terms of human perception it is often associated with brightness of sound.

$$SC = \frac{\sum_{n=0}^{k} n[X(n)]^2}{\sum_{n=0}^{k} [X(n)]^2} \tag{1}$$

3. Spectral Flatness

   It measures tonality of a sound. It gives a low value for noisy sounds and high value for voiced sounds.

$$Spectral flatness = 10 log_{10} \frac{GM}{AM} \tag{2}$$

   GM- Geometric Mean
   AM - Arithmetic Mean

4. Spectral Flux

   Spectral flux measures a change in spectral shape. It is defined as 2-norm of the difference vector between adjacent spectral frame amplitudes.

$$SF_i = || \sum_{n=0}^{k} X_i(n) - X_{i-1}(n) || \tag{3}$$

5. Spectral Rolloff Frequency

   It is a point below which certain amount of power spectral distribution resides.

$$SRF = max_k(\sum_{n=0}^{k} |X(n)|^2 < TH \sum_{n=0}^{k} |X(n)|^2) \tag{4}$$

   TH is the threshold between 0 and 1. We have taken it as 0.85.

6. Spectral Spread

   It is the width of the frequency band of signal frame around centre point of spectrum. It was observed that the value is very high in the noisy part and is comparatively bounded in the part with bird call.

$$BW = \sqrt{\frac{\Sigma_{n=0}^{k}(n - SC)^2 |X(n)|}{\Sigma_{n=0}^{k} |X(n)|^2}} \tag{5}$$

7. Spectral Crest factor

   It gives a measure of peakiness in the signal.

$$SCF = \frac{max(|X(n)|}{\sum\limits_{n=0}^{k} |X(n)|^2} \tag{6}$$

8. Zero Crossig Rate

   It is a number of time-domain zero crossings in a particular frame. A zero crossing occurs when adjacent samples have different signs.

$$ZCR = \sum_{n=0}^{k-1} |sgn(x(n)) - sgn(x(n+1))| \tag{7}$$

   where x is the time-domain signal frame and k is the size of the frame.

9. Intensity

   Intensity was calculated as the amplitude in each frame in decibel scale. The curve was smoothened out further to get the exact number of peaks correctly. Smoothening was done using low pass filter along with spectral spread which was used to

remove peaks due to noisy parts.
intensitynew(j) = α*intensityold(j) + (1-α)*intensitynew(j-1)
intensitynew: filtered output
intensityold: unfiltered output
And thus the intensity curves were classified as flat, single peak, two peaks and multiple peaks.

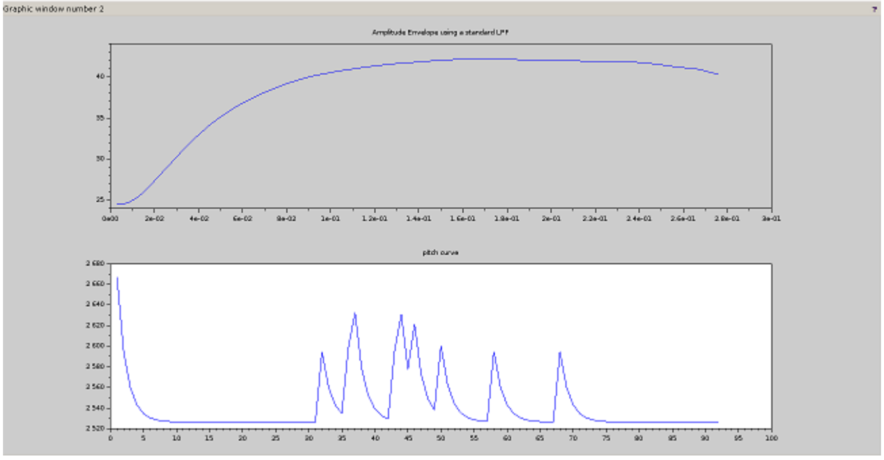


Figure 4: Almost flat intensity curve of Common Lora with a low average pitch

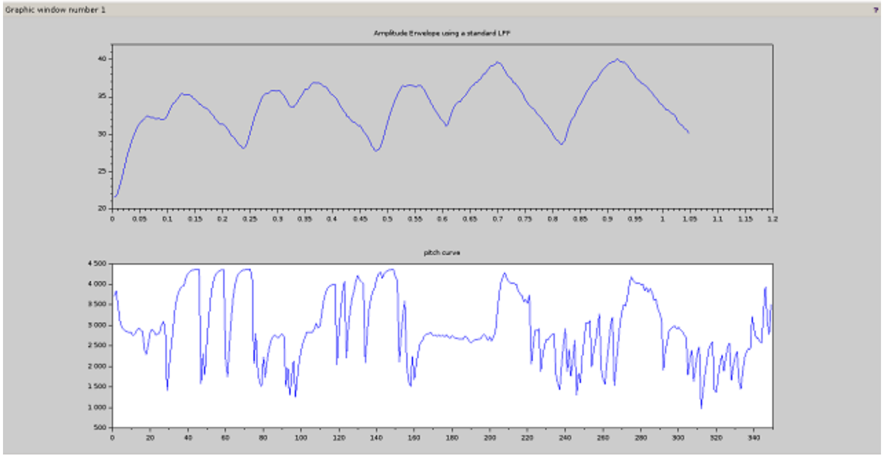


Figure 5: Intensity curve of Long-tailed Shrike with multiple peaks and relative high mean pitch

10. Pitch

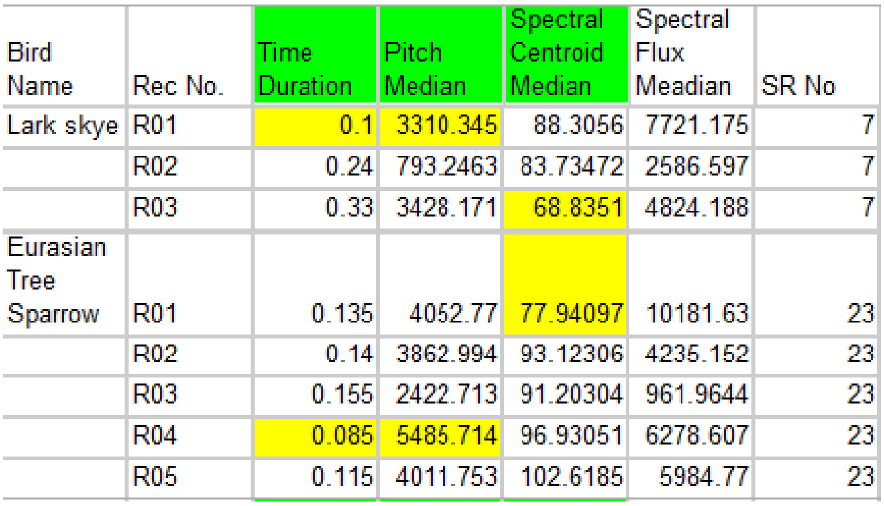
Pitch was calculated using autocorrelation function. The frames of original sound signal were autocorrelated and the inverse of difference between amplitudes two maximum peaks was recorded as pitch for each frame. The curve thus obtained was very peaky but the array was good enough to get good similarities between mean pitch for same bird's samples.

3.3 Conclusions about extracted features

After analysing all the above-mentioned features of the sound signal and the spectogram for different samples of different birds, we came to the following conclusions -

- The feature of intensity class was a discrete one and thus could not be used as a feature vector in the clustering algorithm.
- Amongst rest of the continuous features, time duration, median of spectral flux, spectral centroid and pitch were found to be quite promising. And as a first step we decided to test the clustering algorithm for these 4 features.
- Even though a few accurate features were taken into consideration, we did have a few outliers in out dataset which affected the accuracy of clustering algorithms. The following could be the reasons for the presence of outliers
  1. In case of birds which have the presence of harmonics, there is a possibility that f is taken as pitch at some times and kf (k times f, where k is a natural number) is taken as pitch at others. This problem can be sorted out by taking GCD of the pitch.
  2. Birds have different type of calls and also same call can have more than one type of syllable. So a manual misclassification of two different syllables as the same can lead to an outlier.
  3. Also there could be errors in time duration as the segmentation is manual, so there could be samples where even noisy part is taken into the sample or otherwise samples wherein even the bird-call part has been clipped off.

Here are a few examples of outliers -



Bird Name	Rec No.	Time Duration	Pitch Median	Spectral Centroid Median	Spectral Flux Meadian	SR No
Lark skye	R01	0.1	3310.345	88.3056	7721.175	7
	R02	0.24	793.2463	83.73472	2586.597	7
	R03	0.33	3428.171	68.8351	4824.188	7
Eurasian Tree Sparrow	R01	0.135	4052.77	77.94097	10181.63	23
	R02	0.14	3862.994	93.12306	4235.152	23
	R03	0.155	2422.713	91.20304	961.9644	23
	R04	0.085	5485.714	96.93051	6278.607	23
	R05	0.115	4011.753	102.6185	5984.77	23

Figure 6: Pitch for sample 1 and 3 of Lark syke is almost 5 times that of sample 2. Time duration of sample 4 of Eurasian Tree Sparrow is distinctly different from other 4.

3.4 Clustering and Classification

Initially we tried unsupervised c-means clustering wherein n centers were assigned iteratively until convergence and on each assignment, all the samples were divided in 'n' clusters on the basis of the shortest distance from the cluster. Here, 'n' denotes the number of clusters. Here is the code snippet for k-means clustering:

```
%% Extracting data
[num, txt, raw] = xlsread ('data2.xlsx', 'data without outliers manual');
X=num(:.1:4);
Y=num(:,5);
[Samples, Centres] =kmeans(X,46) ;
```

The major issues with this kind of classification were -

- If 'n' was kept to be equal to number of birds, the outputs were very poor and not as expected.
- There was no sure shot way to decide the number of clusters and also id number of clusters were to be lesser than number of birds, there should be new accurate bases of hierarchial classification which could be used as next rungs of classification. So, this would involve more work and no sureshot accuracy.

So, we shifted to naive-bayes supervised clustering technique. Here, the algorithm assumes a distribution for each cluster (gaussian in our case) and then calculates the parameters for the respective clusters. Then, when a new feature vector is introduced (test case), the probability wrt each custer parameter is calculated. More the probability, more are the chances of belonging to that particular cluster. We used confusion matrix as a measure of accuracy. The most important assumption of this method is that all the features are independent of one another.

$$p_\theta(x,y) = p_\theta(x|y)p_\theta(y) = p_\theta(x_1|y)...p_\theta(x_d|y)p_\theta(y) \tag{8}$$

where x is a d-dimensional feature vector and y is the label ie bird number in our case. Here is the snippet of the code for clustering:

```
%% Extracting data
[num, txt, raw]=xlsread('mydata.xlsx','data without outliers manual');
X=num(:,1:4);
Y=num(:,5);

%% Fit Naive Bayes Model
O = NaiveBayes.fit(X,Y);
C = O.predict(X);
cmat = confusionmat(Y,C);
accuracy=trace(cmat)/sum(sum(cmat))
```

```
count=ones(size(cmat,1),1);
j=1;
number(1)=1;
for i=2:size(Y)
    if Y(i)==Y(i-1)
        number(j)=number(j)+1;
    else
        j=j+1;
        number(j)=1;
    end
end

%%Estimating error
for i=1:size(cmat,1)
    if cmat(i,i)==number(i)
        error(i)=0;
    else
        error(i)=number(i)-cmat(i,i);
    end
end

datapoints=sum(number)
```

Thus, this method is invariant of independent features, and as there is no recurrence or iterative convergence involved, this method is faster.

4 Results

The accuracy of the latter method was much better than the former one. Even in case of Naive Bayes, the accuracy was observed to increase from 88% to 95% by removing a few outliers manually and reducing the dataset of 50 birds to 46 birds. Test samples were formed and following are the results obtained for a few test samples tested on data without outliers-

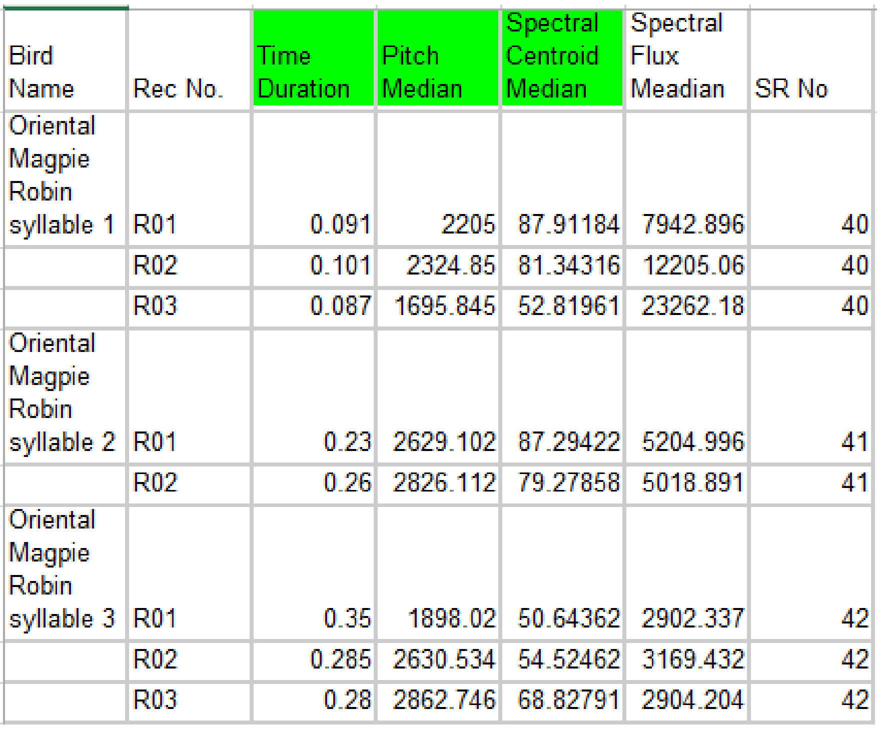
Bird name	Probability of being correct bird	Prediction
Eurasian Tree Sparrow	0.4097	Wrong
Green Heron	0.9909	Correct
Hyacinth Macaw	0.9983	Correct
Oriental White Eye	0.9926	Correct
Red Jungle Fowl	0.9997	Correct

Table 1: Results

5 Problems Faced

There are a few problems which we faced or have been facing -

- The algorithm referred from the paper by Aki Harma, gave the type of harmonic ie which harmonic is the most prominent in that particular signal. But that being a discrete feature, it could not be put into use further.
- Time duration for syllable samples of same birds were quite close, so the naive-bayes clustering algorithm developed a strong gaussian around this feature and thus the prediction on test cases is strongly dependent on the observed time duration.
- There is a possibility that single bird has different types of syllables, so they need to be identified while manual segmentation. Also different clusters need to be formed for different syllables so it results into 2-3 clusters for one bird. One of the examples is shown in figure 7.



Bird Name	Rec No.	Time Duration	Pitch Median	Spectral Centroid Median	Spectral Flux Meadian	SR No
Oriental Magpie Robin syllable 1	R01	0.091	2205	87.91184	7942.896	40
	R02	0.101	2324.85	81.34316	12205.06	40
	R03	0.087	1695.845	52.81961	23262.18	40
Oriental Magpie Robin syllable 2	R01	0.23	2629.102	87.29422	5204.996	41
	R02	0.26	2826.112	79.27858	5018.891	41
Oriental Magpie Robin syllable 3	R01	0.35	1898.02	50.64362	2902.337	42
	R02	0.285	2630.534	54.52462	3169.432	42
	R03	0.28	2862.746	68.82791	2904.204	42

Figure 7: Oriental Magpie Robin has 3 different types of syllables

- The code that we are working on for autonomous segmentation gives syllables and not phrases. So phrase based features cannot be used for the classification. This is not a major problem currently as our clustering is syllable-based, but there are a few phrase based features like time duration between two phrases and number of intensity peaks in a single phrase which could prove to be a good basis of classification. Also more work is required in this section.

- There are a few ouliers in the data. Just because of presence of one outlier the variance of a particular feature for a bird increases significantly thus making it a weak classifier feature. The probable reasons for these outliers are mentioned in 3.3. So, the reason behind their presence needs to be analysed and an algorithm for their removal needs to be designed.

6 Future Work

The work done till now needs a lot more testing and corresponding improvement in order to make it implementable. Future work includes -

1. The reason for occurence of outliers needs to be analysed and rectified accordingly.
2. The number of datapoints in the dataset needs to be improved ie number of samples per bird should be increased to 4-5 on an average.
3. The code for segmentation needs to be completed soon and the complete testing should be done with automatically segmented syllables. Thus making the whole process automatic, once the sound is recorded and feed in the code.
4. All of our present work concentrates only on birds with calls and not songs. The present algorithm needs to be tested for song calls or otherwise a method needs to be devised which would work for song calls.

7 References

- A. Harma, ”Automatic recognition of bird species based on sinusoidal modeling of syllables, in IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP 2003), Hong Kong, April 2003
- Seppo Fagerlund, Automatic Recognition of Bird Species By Their Sounds, Master Thesis, Helsinki University of Technolofy, Department of Electrical and Sommunications Engineering, Nov 8, 2004
- Aki Harma and P. Somervuo, ”Classification of Harmonic Structure in Bird vocalizations”, In proceeding of: Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP ’04). IEEE International Conference
- George, E. B. and Smith, M. J. T. (1997), ”Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model”, IEEE Trans. Speech and Audio Processing 5(5), 389406
- http://www.xeno-canto.org/ for procuring bird sounds
- http://avocet.zoology.msu.edu/ for procuring bird sounds