Automatic bird sound detection in long range field recordings using Mel filter bank features

Project Idea Report

Suhas Bettapalli Nagaraj

Spring 2020

Abstract

The topic of bio-monitoring of fauna, especially that of birds is an ongoing research topic considering the effect of globalization has had around the world. Although huge datasets of bird sound recordings have been collected and annotated over the last decade, the classification of such sounds into bird and non bird sounds has been painstaking work, sometimes requiring manual processing. The goal of the IEEE Research Challenge in 2016 has been to address this concern and to help development automatic algorithms for the detection of bird sounds [1]. Several current methods for audio signal processing such as mel frequency cepstral coefficients (MFCC) coupled with different classifiers have been used in the challenge by participating teams. This course project being in Wavelets, an attempt has been made into comparing and understanding how wavelet based features perform against the current state of the art methods in audio processing in more detail.

1 Introduction

Monitoring of animals and birds is important in this day to understand the effects of urbanization has had on their habitat. Some recent studies [2, 3] list factors such as spatial heterogeneity, habitat fragmentation and intermediate disturbance as major factors influencing the dwindling numbers of fauna around the globe. The use of sound recordings to check for fauna, especially for the problem of bird detection and subsequent monitoring is well suited since birds can more easily be discovered through sound than though visual inspections. Stowell [1] reviewed some of the paradigms and techniques that have been used for bird sound detection over the past few years.

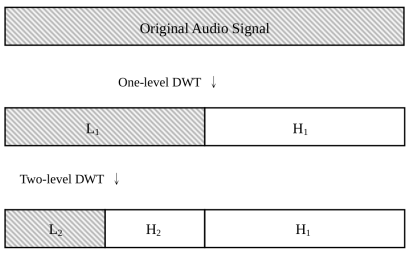
Over the last decade or two, bioacoustics has become one of the most important research areas that has made use of the boom in "big data". One such project by Cornell, called Macaulay Library, has been generating huge amounts of audio, far more than what can feasibly be inspected by humans. The goal of such projects is usually to monitor migration patterns of animal species or to monitor the overall health of the ecosystem. [1, 2] further investigated the utilization of latent acoustic monitoring to evaluate the density of fauna. It is seen that there is renewed interest in trying to work with acoustic records for biodiversity estimations.

Other such massive programmes for the monitoring of birds have preferred using the simpler exists/does not exist characteristic of a particular species in a spatio-transient window [4] instead of working with a single or limited representation of a given species in a particular ecosystem. Rowe observed that automated recognition software improves the perceptability of articulation for different bird species. Through their work, Rowe also observed that with the available technology, manual sifting is needed to set thresholds and certain parameters and also to postprocess the information collected. This could potentially still mean that the total man hours spent on both data collection and subsequently the project would not be reduced when stacked against a manual survey. This shows that while automatic detection methods are quite useful, in practice, automation still needs further development. Rowe and Digby [4, 5] both presumed that upgrades in identification (and characterization) would be preferred, particularly as for alignment and complete automation.

Some of the earliest works combining the fields of audio and wavelets include Tzanetakis [6] who he work looked at three main applications, namely, Speech vs music, identification of Male versus

Feature		Type of transforms	Number of features
Perceptual feature	Subband power P_j	Wavelet	3
	Pitch frequency f_p	Wavelet	1
	Brightness ω_c	Fourier	1
	Bandwidth B	Fourier	1
Frequency cepstral coefficient (FCC) c_n		Fourier	L

(a) Wavelet and Fourier based features [9]



(b) Two level DWT representation [10]

Figure 1: Wavelet representation and some features used in literature

Female voices and identification of Classical music tones using Discrete Wavelet Transform Coefficients (DWTC), MFCC and short term fourier transform coefficients (STFTC). Wavelet based features have been used in other domains such as ECG based Arrhythmia Beat Classification [7] to classifiying percussive sounds [8].

However, wavelet based features have never been utilized in the field of bio-monitoring. It is here that a comparison between previously used methods in the field of bio-monitoring of birds and wavelets can be tested to understand where wavelet based features stand against the current state of the art methods.

Based on literature, a few potential wavelet based features have been discussed below. M. Daniels [8] proposed comparing wavelet based features (db4, db5, and sym5 wavelets) with comparable MFCC features for percussion sound analysis. The dataset for the experiments was collected in-house. The work made use of Support Vector Machines for classification while Lin et al. [9] proposed an audio classification technique which combined wavelets with frequency cepstral coefficients (FCC) as the feature vector. Wavelet features include sub band power and pitch information. The Muscle Fish dataset, which consisted of 410 sounds in 16 different classes to compare different features, is used to evaluate the performance of the features. In a nutshell, the feature consists of Wavelet based features + FCC which is then trained using SVM. A summary of the features used and their respective dimensions are seen in Fig. 1a.

Hsieh et al.[10] proposed a method of extracting wavelet features from audio. The work uses one-dimensional Haar Discrete Wavelet Transform (DWT) to decompose a given frame into three sub bands, namely L2, H2 and H1 respectively. The L2 sub band is then chosen for feature extraction. The two level DWT representation that were used are shown in Fig. 1b. Following the feature extraction, the authors propose using suprasegmental features such as mean, median and standard deviation for training. Lostanlen et al. [11] made use of Mel Spectrograms as their input feature vector with a CNN classifier. The Urban-8K and CLO-43SD datasets were used for classification purposes. The work proposed using both short term data (60mS) & long term recordings (30 mins) for their classification. The work applied a per-channel energy normalization technique in both the time and frequency (TF) domain which achieved an AUC score of 0.6-0.75 based on various threshold settings. T. Pellegrini [12] too conducted experiments using CNNs. The paper proposed using the Mel Filter Bank Energies (MFBE) that are computed on the audio signal as the input feature. The work uses both Freefield1010 and Warblr (FFW 1) datasets. The best performing model had an AUC score of 88.2%. This can be considered as the State of the art method to compare our experimental results with.

2 The Problem

The IEEE Research Challenge in 2016 was on the development of fully automatic algorithms for bird sound detection [1]. While current methods for audio signal processing such as MFCC and Spectrograms have performed well with classifiers, the question that we are interested in is whether wavelet based features hold their own against the current state of the art methods? Can wavelet features contain good discriminating information that other features do not? These questions will be answered at the end of the project.

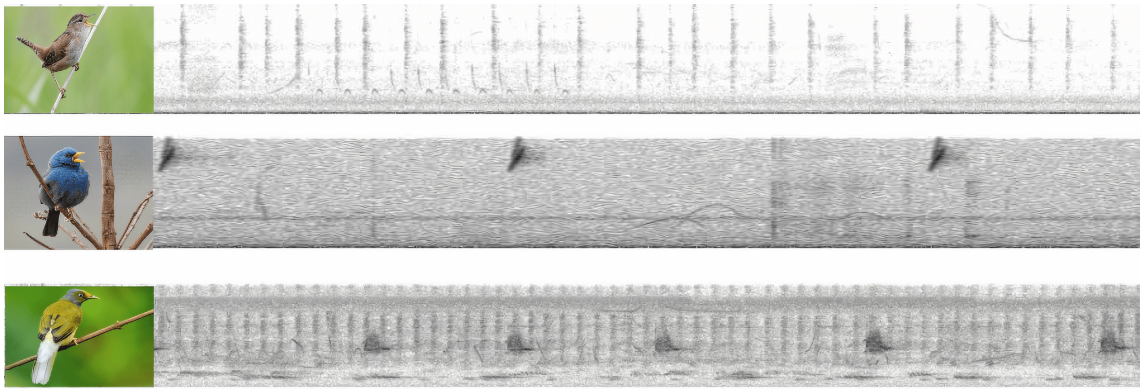


Figure 2: Different birds and their characteristic sound recordings.
Top: **Marsh Wren**, Middle: **Blue Finch**, Bottom: **Gray headed bulbul**

3 Dataset

Long Range Field recordings (freefield1010)

For this course project, the **freefield1010** long range field recording dataset is being used. The dataset consists of ten second recordings of various species of birds. The recordings have been annotated with a hasBird/noBird label to depict the presence/absence of bird sound in each recording. Over seven thousand such field recordings from around the world exist, which have been assembled in this dataset. The recordings are diverse in terms of environment and the locations where the audio has been recorded and thus helps in generalizing results obtained. Some representative examples of birds and their bird sound recordings are plotted over a duration of ten seconds in Fig. 2.

4 Features, methods and evaluation

4.1 Features

Based on the literature review, it is decided that four features, namely MFCC [6, 9], Mel filter bank energies (MFBE)[12], Two level DWT (TDWT) and Mel Spectrogram features (MSGF) [11] would be used for classification purposes. In the review, MFCC and MSGF have been widely used in audio processing experiments. The wavelet based features being used in the project are db4 and sym5 wavelets [8, 10]

To compute the filter bank features, the audio signal is passed through a pre-emphasis filter. It is then sliced into overlapping frames. This is followed by applying a window function to each frame. In each resulting frame, a Fourier transform is applied (specifically a Short-Time Fourier Transform). The power spectrum is then calculated, followed by computing the filter banks. To obtain the MFCC features, the discrete cosine transform is applied to the filter banks thereby retaining a number of the resulting coefficients. The first co-efficient (which contains energy) for example, is discarded. The following step for both filter banks and MFCCs is CMVN (cepstral mean variance normalization) [13].

- Mel frequency cepstral coefficients (MFCC)
- Mel filter bank energies (MFBE)
- Two level DWT (TDWT)
- Mel Spectrogram features (MSGF)

4.2 Methods

Table 1 was constructed based on the literature survey which summarizes some of the methods that could be used along with their advantages/disadvantages. The Voice Activity Detection (VAD) method is generally used for speech signal processing although there have been attempts where it

Method	Algorithms used	Pros	Cons
Presence and onset	Classifiers	Manual annotations can be efficient; overlapping windows can be used	Low temporal precision
VAD based	VAD / HMM	Could help sort out bird sound segments for easier classification	Overlapping events are merged
TF Axis	Spectrogram correlation, pitch trackers	Can classify better than presence and onset based methods through temporal details	Harmonic stacks may separate; sometimes could be inappropriate for non-tonal sounds

Table 1: Possible methods that can be used for classifiying hasBird / noBird sounds

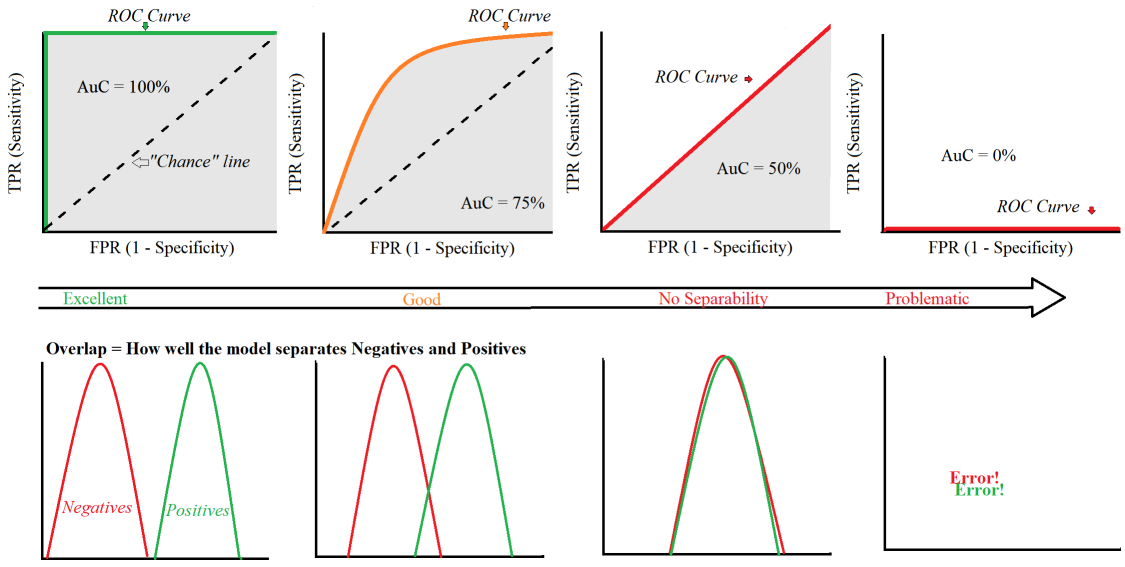


Figure 3: AUC-ROC curves for 2 class classification with True Positive Rate (TPR) vs False Positive Rate (FPR) on the Y & X axis respectively

has been used for non speech sounds with varying degrees of success. Spectrogram-based methods that employ the Time Frequency (TF) axis have also been used in the current State of the Art (SoA) methods.

In this project, two classification methods will be used. They are:

- Support Vector Machines (SVM)
- Convolution Neural Networks (CNN) (SoA)

A block diagram that depicts the project flow is shown in Fig. 4. The dimensions of each feature have been described above the arrow in each case where N is the number of audio segments in each recording.

4.3 Evaluation

- AUC-ROC Characteristic curves : The potential of a network to accurately classify different classes is evaluated via the area under receiver operating characteristic curve. AUC-ROC is a performance measure for classification at different threshold settings. ROC is a probability curve while the AUC represents a measure of separability. [14]. A model with an AUC close to 1(0) reflects a good(poor) measure of separability between the classes. Thus, the higher the AUC, the better is the model at distinguishing between the two classes.

  Fig.3 shows different values of AUC and their interpretation. For an AUC of 1 (or 100%), the model is able to distinguish between the two classes. As the value drops and reaches AUC = 0.5(50%), it means that the model is unable to distinguish between the classes and can be seen as an overlap between the two classes.

5 Challenges involved

- Will the trained model be robust to strongly varying noise such as wind, rain and other fauna?
- Will the methods work only for specific species of birds or is it generic?

6 Acknowledgements

I would like to thank Prof. Vishal Monga for providing a practical course project where I have been able to learn and use wavelet based features for audio classification.

References

[1] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in Audio: A survey and a challenge," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2016.

[2] M. L. McKinney, "Effects of urbanization on species richness: A review of plants and animals," *Urban Ecosystems*, vol. 11, no. 2, pp. 161–176, 2008.

[3] S. Zhang, M. Suo, S. Liu, and W. Liang, "Do major roads reduce gene flow in urban bird populations?," *PloS one*, vol. 8, no. 10, 2013.

[4] K. Rowe, "Automated recognition software improves detectability for a range of bird species' vocalizations," in *Int Bioacoustics Congress (IBAC)*, 2015.

[5] A. Digby, M. Towsey, B. D. Bell, and P. D. Teal, "A practical comparison of manual and autonomous methods for acoustic monitoring," *Methods in Ecology and Evolution*, vol. 4, no. 7, pp. 675–683, 2013.

[6] G. Tzanetakis, G. Essl, and P. Cook, "Audio analysis using the discrete wavelet transform," in *Proc. Conf. in Acoustics and Music Theory Applications*, vol. 66, 2001.

[7] Q. Qin, J. Li, L. Zhang, Y. Yue, and C. Liu, "Combining low-dimensional wavelet features and support vector machine for arrhythmia beat classification," *Scientific reports*, vol. 7, no. 1, pp. 1–12, 2017.

[8] M. Daniels, "Classification of Percussive Sounds Using Wavelet-Based Features," *CCRMA, Stanford University thesis*, 2010.

[9] C.-C. Lin, S.-H. Chen, T.-K. Truong, and Y. Chang, "Audio classification and categorization based on wavelets and support vector machine," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 644–651, 2005.

[10] S.-L. Hsieh and H.-C. Wang, "Feature extraction for audio fingerprinting using wavelet transform," in *National Computer Conference*, 2005.

[11] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Robust sound event detection in bioacoustic sensor networks," *PloS one*, vol. 14, no. 10, 2019.

[12] T. Pellegrini, "Densely connected cnns for bird audio detection," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1734–1738, IEEE, 2017.

[13] H. Fayek, "Speech processing for machine learning: Filter banks, mel-frequency cepstral coefficients (MFCCs) and what's in-between," Apr 2016.

[14] S. Narkhede, "Understanding AUC - ROC Curve," May 2019.

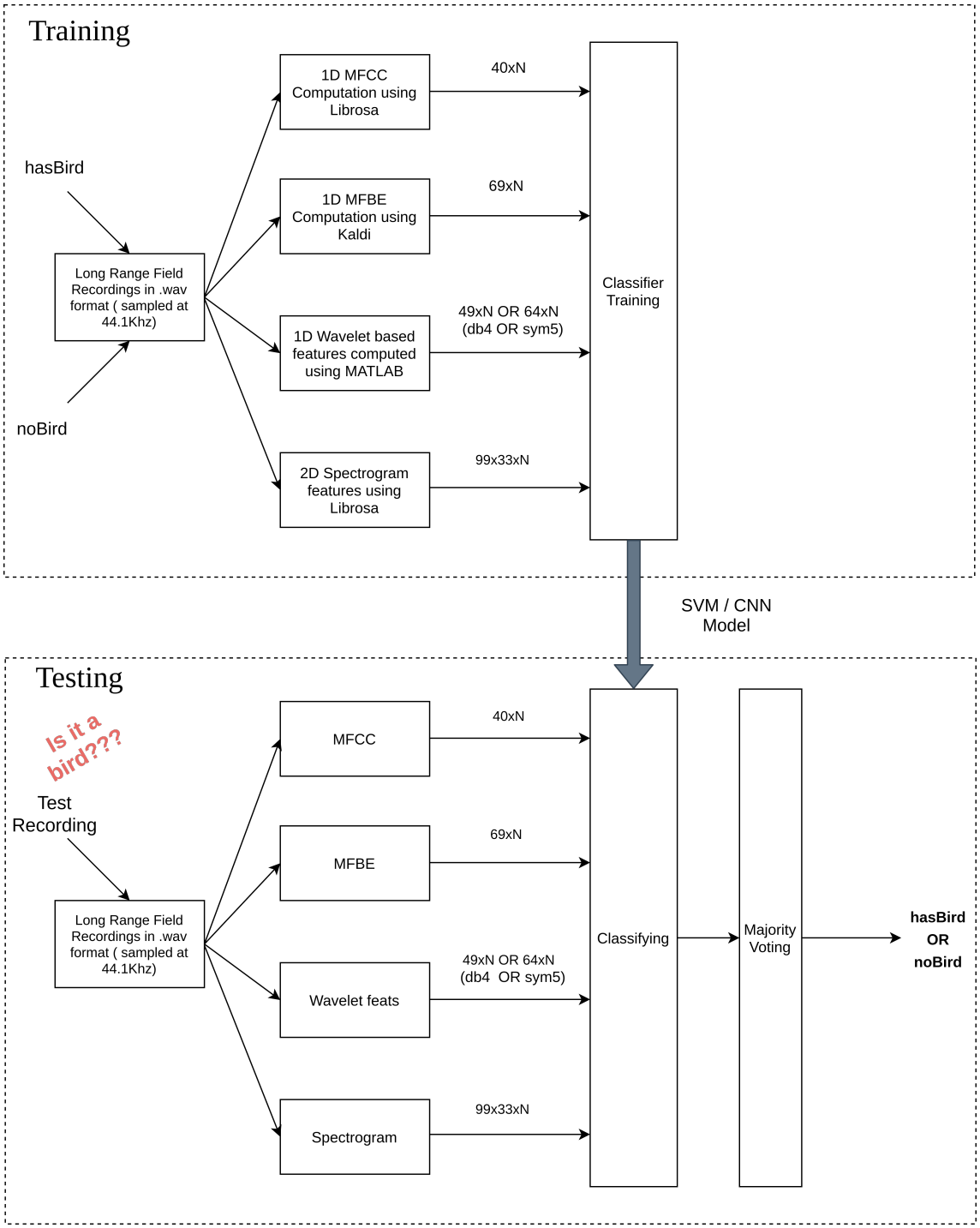


Figure 4: Project outline