MINI PROJECT II

(2019-20)

**AUDIO BIRD DETECTION**

**(Machine Learning)**

**SYNOPSIS**

**Department of Computer Engineering & Applications**

**Institute of Engineering & Technology**





**Submitted to :**

**Submitted by :**

Mr.Amir Khan

Naman Mangal(171500198)

(Technical Trainer)

Kshitiz Gupta(171500166)

Jayendra Singh(171500147)

Humera Shaheen(171500142)

**AUDIO BIRD DETECTION**

**ABSTRACT**

Bird sounds possess distinctive spectral structure which may

exhibit small shifts in spectrum depending on the bird species

and environmental conditions.

Bird sounds possess distinctive spectral structure which may

exhibit small shifts in spectrum depending on the bird species

and environmental conditions.

Bird sounds possess distinctive spectral structure which may

exhibit small shifts in spectrum depending on the bird species

and environmental conditions.

Bird sounds possess distinctive spectral structure which may

exhibit small shifts in spectrum depending on the bird species

and environmental conditions.

Bird sounds possess distinctive spectral structure which may

exhibit small shifts in spectrum depending on the bird species

and environmental conditions.

Assessing the presence and abundance of birds is important for monitoring specific species as well as overall ecosystem health. Many birds are most readily detected by their sounds, and thus passive acoustic monitoring is highly appropriate. Bird sounds possess distinctive spectral structure which may exhibit small shifts in spectrum depending on the bird species and environmental condition.

In this paper, a SVM based linear (or non-linear) classifying bird sounds is presented and tested through different configurations and hyper parameters. This model is fine-tuned using a dataset acquired from the freefield1010 Stowell and Plumbley (2014b) selected from the free sound online audio archive. This method achieves 88.5% Area under ROC Curve (AUC) score on the unseen evaluation data. Gaussian Mixture Model is able to attain performance of around 88% AUC (area under the ROC curve), much higher performance than previous general-purpose methods.

Index Terms— Bird audio detection, Support Vector Machine (SVM), Gaussian Mixture Model

**INTRODUCTION**

Bird audio detection (BAD) is defined as identifying the presence of bird sounds in a given audio recording. In many conventional, remote wildlife-monitoring projects, the monitoring/detection process is not fully automated and requires heavy manual labour to label the obtained data (e.g. by employing video or audio) [1, 2]. In certain cases such as dense forests and low illumination, automated detection of birds in wildlife can be more effective through their sounds compared to visual cues. The problem is challenging as the bird sounds may vary drastically in different species of birds. Automation of this would save lots of manual efforts and makes bioacoustics easier. In a set of audio samples collected in wild, we propose to classify if the audio sample contain bird sounds available in it or not. . This indicates the need for automated BAD systems in various aspects of biological monitoring. For instance, it can be applied in the automatic monitoring of biodiversity, migration patterns, and bird population densities. Using an automated BAD system as pre-processing/filtering step to determine the bird presence would be beneficial especially for remote acoustic monitoring projects, where large amount of audio data is employed. Our work is mainly based on vocal sound made by birds in the audio. Bird calls are often short and serve a particular function such as alarming or keeping the flock in contact. We use spectrograms, which are a visual representation of the magnitude returned by the Short Time Fourier Transform (STFT). STFT is a version of the Discrete Fourier Transform (DFT), which instead of only performing one DFT on a longer signal, splits the signal into partially overlapping chunks and performs the DFT on each using a sliding window. MFCC based features were extracted for each recording for the pre-processing of dataset. Further training and testing of data was performed using SVM. The main contribution being the combined approach of MFCCs and SVMs classification. Such classification technique for the modelling of SVMs to perform audio classification with lower error rates is presented in this paper. Advantages of both resulted in increased accuracy and faster classification method. Basic objective was to decrease error rate with simplified computations.

**RELATED WORK**

Extensive research has been conducted in the recent past dissecting potential approaches to the presented issue. A rise in interest may be attributed to the annual Bird CLEF [6] recognition challenge: a biodiversity data evaluation campaign. The training dataset of Bird CLEF 2017 comprises over 36,000 audio files from 1500 different species, collected from Xenocanto, with classes not necessarily having an equal number of sound samples. The challenge focuses on recognizing single audible species as well as separating multiple overlayed sounds in field recordings. In 2017, Kahl et al. [7] undertake the challenge, with their experiments measuring 60% accuracy for overlayed sounds and 68% for recognizing the dominant species. In 2016, Piczak [8] approaches this problem similarly with 41.2% accuracy for multi-labelled and 52.9% for single labelled data. His experiments involve networks trained from scratch, Mel scaled power spectrograms, an upper frequency cap and noise filtering. From the whole literature, we find that CNN’s and RNN’s are proven to have good accuracies and perform better. Their efficiencies are nearly about 80%. Most of the contenders for the Bird Audio detection uses the CNN classifier with elaborative pre-processing technique.

**DATA SET AND EVALUATIONS**

We are using two available datasets i.e. field recordings dataset and crowd-sourced dataset Field recordings contain over 7000 samples recorded around the world. The crowd sourced dataset contains over 10,000 ten-sec audio clips recorded using mobile phones over wide locations in the UK. The datasets are binary labelled, denoting presence or absence of bird sound.

For now, field recording data is used as the data set in which is divided into Training and Test set in 90:10 ratio. Before training our data through different classifier, we will pre-process it both in time domain and frequency domain (even in cepstral domain!) and extract the Mel Frequency Cepstral Coefficients (MFCC). These coefficients are the optimized features.

**METHODOLOGY**

Step 1: The first step is to apply a pre-emphasis filter on the signal to amplify the high frequencies. A pre-emphasis filter is useful in several ways: (1) balance the frequency spectrum since high frequencies usually have smaller magnitudes compared to lower frequencies, (2) avoid numerical problems during the Fourier transform operation and (3) may also improve the Signal-to-Noise Ratio (SNR).

The pre-emphasis filter can be applied to a signal xx using the first order filter in the following equation:

y(t)=x(t)−αx(t−1)

Step 2: After pre-emphasis, we need to split the signal into short-time frames. We can safely assume that frequencies in a signal are stationary over a very short period of time. Therefore, by doing a Fourier transform over this short-time frame, we can obtain a good approximation of the frequency contours of the signal by concatenating adjacent frames.

Step 3: After slicing the signal into frames, we apply a window function such as the Hamming window to each frame. We can now do an NN-point FFT on each frame to calculate the frequency spectrum, which is also called Short-Time Fourier-Transform (STFT), where NN is typically 256 or 512, and then compute the power spectrum (periodogram).

Step 4: The final step to computing filter banks is applying triangular filters, typically 40 filters, on a Mel-scale to the power spectrum to extract frequency bands. The Mel-scale aims to mimic the non-linear human ear perception of sound, by being more discriminative at lower frequencies and less discriminative at higher frequencies.

Step 5: filter bank coefficients computed in the previous step are highly correlated, which could be problematic in some machine learning algorithms. Therefore, we can apply Discrete Cosine Transform (DCT) to decorrelate the filter bank coefficients and yield a compressed representation of the filter banks. Typically, for Automatic Speech Recognition (ASR), the resulting cepstral coefficients 2-13 are retained and the rest are discarded. Finally, applied the de-emphasis filter. Finally, normalise the MFCC coefficients.

**EVALUATION**

We will validate our results through evaluation of following parameters:

* TSNE plot for class separability
* ROC Curve
* Classification Accuracies
* Confusion Matrix

**FUTURE WORK**

So far, we have trained our data with SVM classifier and we will proceed towards making this accuracy better. Either we train our model with some different data set or apply changes in feature extraction techniques.

Finally we will train the data with rest of the two classifiers.

There are lots of option for the Data sets available on the net. So we can test our results on that.

There will be no amendments in the evaluation metrics further.

We will also try to train our data through Convolutional Neural Nets (CNN)

Individual contribution will be as follows:

Gaurav Singh: Gaussian Mixture Model

Abhishek Gupta: Random Forest Algorithm

Ashutosh Vaish: Study of CNN and improve SVM results

**ANALYSIS AND PROGRESS:**

1: While observing dataset closely, we found that positive labels are about 1/3rd of the whole data set. Due to this Imbalance, it reflects in our training and testing accuracies.

2: Data set is collected in wild due to which it is difficult to restrict the frequency band responsible for bird sounds. There are sounds of other natural activities like sound of flowing water, similar sounds of other animals that resembles the bird chirp.

3: In our data, sound of birds is for only the fraction of time. Classification is quite difficult.

4: We have chosen spectrogram as our feature extraction technique from which we have extracted MFCC coefficients. Each frame have certain number of coefficients. So whole sample have large number of MFCC coefficients in vectored form. Temporal Information of data is lost.

5: Current work uses MFCC coefficients directly to train SVM model without any dimension reduction technique. Dimension reduction technique like PCA, GPCA (Gaussian PCA), KPCA could result in better results.

6: Samples in the data set have uneven distribution of energy which we can see in the power spectrogram. Energy normalisation is necessary.

7: In time domain, we have used pre-emphasis filter to balance the frequency spectrum since high frequencies usually have smaller magnitudes compared to lower frequencies.