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

Suhas BN
EE Department
The Pennsylvania State University
University Park, USA 16802
bnsuhas96@gmail.com

***Abstract*—The topic of bio-monitoring of fauna, especially that of birds, is an ongoing research topic. Although huge datasets of bird sound recordings are available, the classification of such sounds into bird & 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 develop automatic algorithms for the detection of bird sounds [1]. An attempt has been made to compare and understand how wavelet based features perform against the current state of the art methods in audio processing in more detail. Two statistical/deep learning approaches namely Support Vector Machines and Convolutional Neural Networks have been used to compare the results. Since most of the works in audio signal processing have employed Fourier based methods, there has been a stagnation in the development and usage of newer/other features for experiments. This is one of the points that has been addressed in this work. The Receiver Operating Characteristic curve (ROC) has been employed to test the diagnostic ability of the audio features. Wavelets have performed consistently and at a level similar to the best performing Fourier based methods highlighting the possibility of using such features as a viable alternate for future audio processing experiments.**

***Keywords*—wavelets, audio processing, filterbanks, bird sound detection**

I. Introduction

Monitoring of animals and birds is important in this day to understand the effects 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 through 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. This has seen 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 perceptibility 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 number of 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.

II. Related Work

Some of the earliest works combining the fields of audio and wavelets include Tzanetakis [6] where the authors looked at three main applications, namely, Speech vs music, identification of Male versus 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 classifying percussive sounds [8].

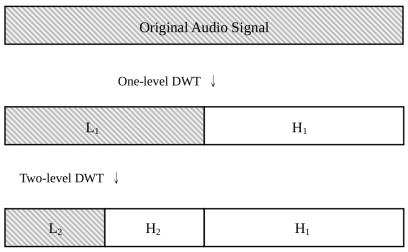
However, wavelet based features have never been utilized in the field of bio-monitoring. It is here that a comparison between previously used methods in the field of bio-monitoring of birds and wavelets can be tested to understand where wavelet based features stand against the current state of the art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

Based on the 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 can be 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 representations that were used are shown in Fig. 1b. Following the feature extraction, the authors use 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.

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

III. 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, MFBE and log-scaled Mel 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? These questions will be clear at the end of this work.

IV. Dataset

Long Range Field recordings (freefield1010)

For the experiments, 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.

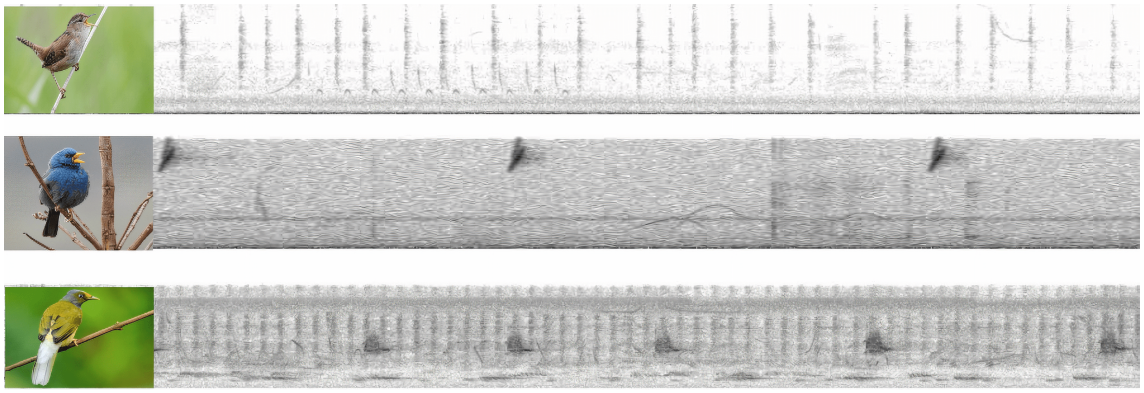


Figure 2: Different birds and their characteristic sound recordings.
Top: **Marsh Wren**, Middle: **Blue Finch**, Bottom: **Gray headed bulbul**

V. Why Wavelets?

The wavelet transform (WT) has applications in both stationary and non-stationary signals such as removal of electrical noise, detecting abrupt changes and data compression.

Using WT, we can decompose a signal into a group of smaller signals, called wavelets. Each of these wavelets are well defined and just like the Fourier transform (FT), have a dominant frequency.

In WT, the wavelets are short interval or transient functions centered around a specific time. The issue with the FT arises when shifting from time to frequency domain, and some important information about time variations is lost.

Unlike FT, WT allows us to analyze time & frequency together. This provides information on how the frequency content evolves over time. WT when discretized is called the Discrete Wavelet Transform (DWT). The **main advantage** that a wavelet offers is that it decomposes the audio into many scales which represent a wide range of frequencies, and, for every scale, the position of the WT can be found at any important time point. At these points, we can eliminate any noise effectively. Short-term wavelets allow information to be extracted from the High Frequency (HF) components since electrical noise is more likely to exhibit HF fluctuations [13], [14]. Long term wavelets allow us to obtain information from LF. With the information of the HF and LF components, we can then specify a threshold [15].

Generally, noise is a rapidly changing signal. This quick change implies that it consists of high frequency components which after decomposition can be removed.

VI. Features, methods and evaluation

A. *Features*

Based on the literature review, it was initially decided in the project idea that four features, namely MFCC [6], [9], Mel filter bank energies (MFBE)[12], Discrete Wavelet Transform features namely, Daubechies wavelet 4 (db4) and Symlet 5 (sym5) [8], [10] and Mel Spectrogram features (Mel Spec/SPEC) [11] would be used for classification purposes. It is to be noted that most of the papers in the IEEE Challenge have extensively made use of MFCC and Mel Spec and their subsequent suprasegmental features in the audio processing experiments.

However, it is now decided to expand the scope of experiments just to understand the ability to distinguish between hasBird/noBird features among different audio features (especially within wavelets). The Wavelet feature set has been expanded to include the following : db 3,4,5 and sym 4,5,6.

Along with this, the features used in four reference papers have been made use of for this paper. These four papers would thereafter be referred to as Paper 1 (p1) [10], Paper 2 (p2) [6], Paper 3 (p3) [11] and Paper 4 (p4) [9].

The audio features have been grouped into three groups : Purely Fourier, Purely Wavelet based and Mixed which comprises of both Fourier and/or Wavelets and are based on reference papers.

To simplify the procedures for extraction and also to avoid any computational variations among features in the same group (for example, in the Fourier group, MFCC from Librosa and MFBE from Kaldi may have different resolutions for use in calculations and hence may affect the results obtained), it was decided to extract all the features from MATLAB. Some of the features that do not have built-in functions have been coded from scratch.

The original signal is filtered through a high pass filter (HPF) g(n) and a low pass filter (LPF) h(n). This is followed by downsampling to obtain the decomposed signal through both HPF and LPF. This happens to be half the length of the original signal. This results in the decomposition of the audio signal into distinct frequency components. The LF components are called approximations and HF components are called details. This constitutes one

level of decomposition, mathematically expressed as follows :

$$Y_{hp}(k) = \sum_n X(n)g(2k-n)$$
$$Y_{lp}(k) = \sum_n X(n)h(2k-n) \quad \text{(VI-A.1)}$$

where $X(n)$ is the original audio signal, $h[n]$ and $g[n]$ are the impulse responses while Y_{hp} and Y_{lp} are the outputs of the HPF and LPF respectively, after subsampling by a factor of 2. This procedure is known as sub-band coding. An example for an audio signal with details and approximations (along columns) for different decomposition levels (along rows) of the daubechies wavelet has been shown in Fig.3.

To extract the wavelet features, *wavedec* function was used on MATLAB. Following this, the *detcoef* function is used on MATLAB to obtain the 1-D detail coefficients at level N = 6. Level 6 was chosen after comparing between the size of the feature vector and the relative performance it provided and N = 6 was found to perform well without loss of accuracy. The extracted feature vector has an order much higher than the original signal. For this reason, the extracted feature was subsampled to reduce it to a more managable form.

It is to be observed that subsampling did not affect the performance of the feature since it does not affect the resolution. In the case of subsampling by a factor of two, removing half of the spectral components present in the signal will make half the total number of samples superfluous and thus can be left out without any loss of information.

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 Filter banks, we apply triangular filters on a Mel-scale to the power spectrum to extract frequency bands. This aims to mimic the non-linear human perception of sound (i.e. good discrimination at lower frequencies & less so at higher frequencies).

To convert between Hertz and mel:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$
$$f = 700\left(10^{m/2595} - 1\right) \quad \text{(VI-A.2)}$$

Each filter in the triangular filterbank can be modelled as follows:

Group	Features
Wavelet based	db3 (41), db4 (54), db5(45) sym4 (54), sym5 (54), sym6 (54)
Fourier based	MFCC (55), MFBE (51), Mel Spec (64)
Mixed	Paper1 (p1) (43) Paper2 (p2) (43) Paper3 (p3) (40) Paper4 (p4) (41)

Table I: Audio features and their dimensions in brackets used in this work

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \le k < f(m) \\ 1 & k = f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \quad \text{(VI-A.3)}$$

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) [16].

Some of the features that have been used in previous works being implemented in this paper include the following:

- **Energy of a signal** $x(n)$ decomposed into approximations a_n and details d_n at a particular scale m:

$$\sum_{n=1}^{N} \left|x(n)^2\right| = \sum_{n=1}^{N} |a_n^m|^2 + \sum_{m=1}^{M}\sum_{n=1}^{N} |d_n^m|^2 \quad \text{(VI-A.4)}$$

- **Wavelet variance**, which is a scale-by-scale decomposition of variance of signal at a particular scale m :

$$\left\langle T_{m,n}^2 \right\rangle_m = \sum_{n=0}^{2} \frac{(T_{m,n})^2}{2^{M-m}} \quad \text{(VI-A.5)}$$

Here, $T_{m,n}$ represents the DWT coefficients.

- **Fluctuation intensity (FI)** measures the energy distribution across different scales m :

$$FI = \frac{\left[\left\langle T_{m,n}^4 \right\rangle_m - \left(\left\langle T_{m,n}^2 \right\rangle_m\right)^2\right]^{1/2}}{\left\langle T_{m,n}^2 \right\rangle_m} \quad \text{(VI-A.6)}$$

The original freefield1010 dataset has been used to compare between three groups of features as listed in Table I.

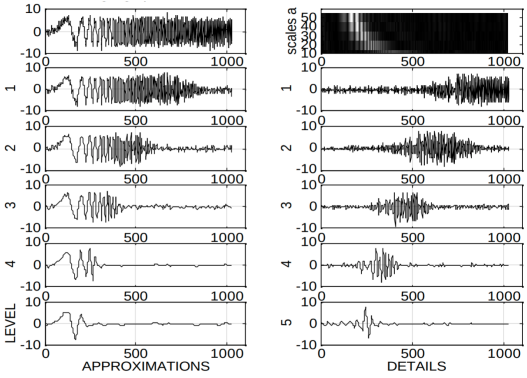


Figure 3: The DWT of a 1 second audio signal using the daubechies 6 wavelet. Each row represents a different level and the columns represent the Approximation and Details respectively

B. Methods

Table II 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 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 methods. In this work, two classification methods will be used. They are:

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

While CNN has been considered a state of the art method owing to the advances in the field of deep learning and the use of visualizing audio through log-scaled Mel spectrograms, traditional audio features trained on SVM still perform well as seen in the results section below.

A block diagram that depicts the work flow is shown in Fig. 5. The dimensions of each feature have been provided in brackets.

C. 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. [17].

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.4 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.

VII. Challenges involved

Although the **freefield1010** dataset has been annotated, the field recordings data still contains a lot of noise due to environmental/man-made factors (such as vehicles).

Noise covers a broad spectrum of frequencies. Field recording generally has noises and this might create two significant problems. First, since it is broadband, it has the potential to overlap sounds that may provide vital information.

Secondly, since it has the potential to overlap other sounds, it becomes tedious to remove the noise completely without taking out the good sound features along with it. Given a scenario where we depend on the Machine Learning model to identify a bird's sound in real time, we can encounter a situation with other noisy constituents around us. It is here that it would be interesting to find out how each of the features perform and if they are robust enough to take care of strongly varying noise seen in nature such as wind, rain and other fauna?

In this work, although the audio does have some field recording noise, we still try to add some Gaussian noise to the original audio signal to mimic the same and to observe the robustness of the audio feature. This has been discussed in the following section. It must be added here that a more robust dataset with recordings from microphone arrays can unlock more from the data than we currently have. This can advance the study of not just flora and fauna but so much more. One of the **benefits** of using a wavelet based method is that it can be used to denoise the audio signal using a thresholding technique [13], [15]. The performance of wavelet based features (especially with AWGN) will be discussed in the following sections.

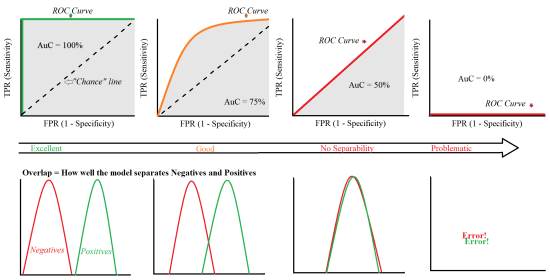


Figure 4: The AUC-ROC curves for 2 class hasBird/noBird classification with True Positive Rate (TPR) vs False Positive Rate (FPR) on the Y & X axis respectively

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

VIII. Experiments

While the original plan has been to compare between four features, namely MFCC, MFBE, TDWT and SPEC, the experimental setup has been further expanded to include the following:

- Along with the discussed db4 and sym5 wavelets, we can compare between similar TDWT features such as db3, db5, sym4 and sym6.
- To understand if the audio features are robust to strongly varying noise such as wind, rain and other fauna, the original audio is added with White Gaussian Noise to mimic natural conditions. This has been discussed in further detail below.

The same set of experiments have been repeated for the audios that have been corrupted with Additive White Gaussian Noise (AWGN). AWGN is a type of noise that has been used substantially in the field of information theory to simulate the effect of random processes seen in nature. This work being on long range field recordings, an effort has been made to understand the robustness of the audio features to four varying degrees of AWGN namely, 0.2, 0.4, 0.6 and 0.8. In this regard, four sets of experiments have been performed with adding AWGN to the audio data. The original audio signal is normalized. For all the experiments in this work, a ten fold cross validation setup was used. Ten groups each with 4000 samples of 10 second duration of hasBird and noBird recordings have been used. Nine groups have been used for training and the remaining one group for testing in a round robin fashion. The SVM training has been done on MATLAB with the Medium Gaussian SVM and Quadratic SVM performing the best among the chosen ones (depicted for each feature above their respective SVM-ROC plot - see Section IX). For the CNN, (network architecture in Fig. 6) the training has been done using Keras [18] with Tensorflow [19] in the backend.

The input feature SPEC has dimension of 96×33 with Melbins = **96**, and an audio length of 10 second represented by **33** frames. The number of convolutional filters or 'feature maps' is **32**. The dimensions at the top of each block represent the output feature dimension after that layer. The activation function used is ReLU (except for softmax before output). A 3×3 convolutional kernel (represented by 2D Conv) is used. Every convolution layer of size $h \times w \times d$ learns ' d ' features of size $h \times w$, where h and w refer to the height and width of the kernels learnt. The size of pooling area for max pooling (represented by Max Pool) is 2×2. The optimum convolutional layer dropout is experimentally set to **0.5** while the dense layer dropout is set to **0.6**. The output at the final dense layer is **1** (hasBird or noBird). The training is done using Categorical cross-entropy as the loss function with Adadelta optimizer working the best among the chosen ones.

IX. Results

Fig. 7,8 and 9 contains the t-SNE and their corresponding AUC-ROC plots for Wavelet based, Previous works and Fourier based methods using the SVM classifier respectively. These three figures provide us details such as how well the two classes can be distinguished visually and also about how well a feature can help in identifying bird sounds better.

Although the t-SNE plots do not visually offer distinct separation between the two classes, we can still see one of the classes being present predominantly in one direction (e.g. horizontal/vertical, etc.) This characteristic feature could be a means in better identifying the correct class during classification. However, in the case of MFCC features, the t-SNE plot gives a good set of clusters for each of the two classes. This argument is well supported by the AUC-ROC score of MFCC (1.0).

In classification tasks, the wavelet based features namely, db 3,4,5 and sym 4,5,6 consistently perform at a range of ROC = 0.96 to 0.99. Paper 1 performs at a slightly lower level with ROC = 0.84 while Papers 2 through 4 perform well at ROC 0.98-0.99 using SVM.

While the Fourier based features are the worst performing among the lot with SPEC =0.75 and MFBE at a close ROC = 0.76, MFCC based features is able to achieve a perfect ROC = 1.0.

The above experiment is repeated using the CNN classifier. The AUC-ROC curves are shown in Fig.10, 11 and 12 respectively. The wavelet

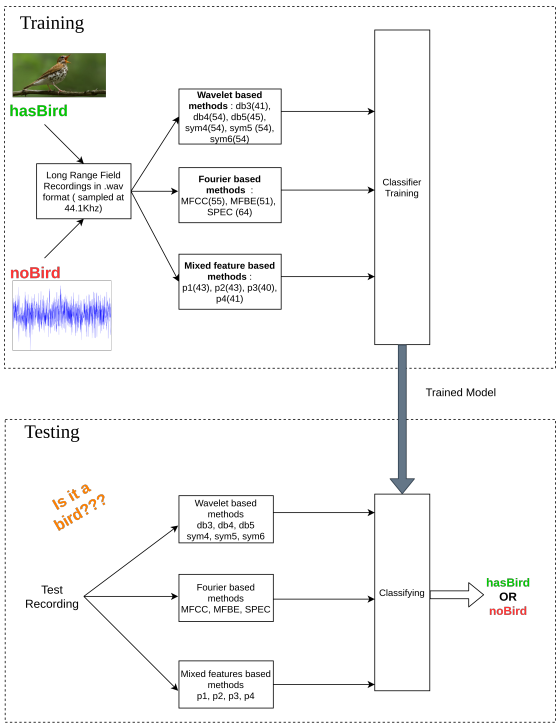


Figure 5: Brief Outline

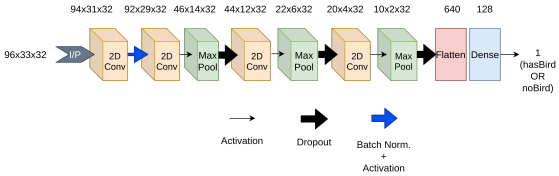


Figure 6: Illustration of 2D CNN architecture proposed for classification.

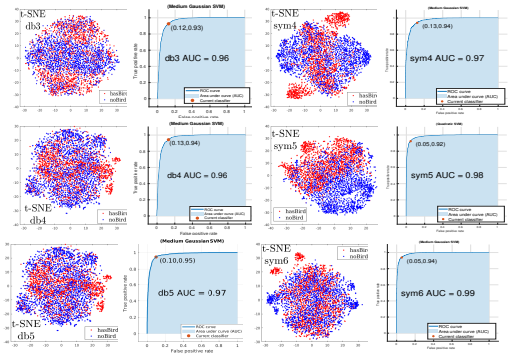


Figure 7: t-SNE plots and SVM based AUC-ROC curves for 2 class bird sound classification with wavelet based features

features were observed to perform well across classifiers and the results are consistent. Considering the features used in previous works (paper1 to paper4), it is observed that while Paper2 shows a small degradation in the performance, Paper4 performs consistently across classifiers. However, it is observed that while Paper1 shows a slight

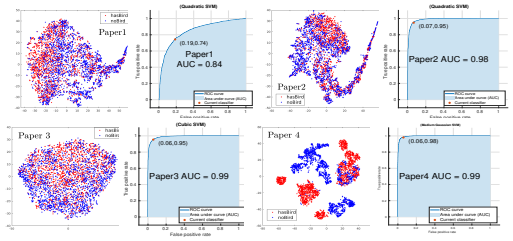


Figure 8: t-SNE plots and SVM based AUC-ROC curves for 2 class bird sound classification with features used in previous works

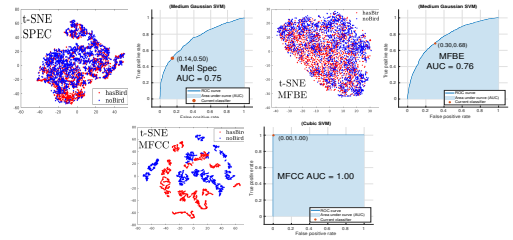


Figure 9: t-SNE plots and SVM based AUC-ROC curves for 2 class bird sound classification with fourier based features

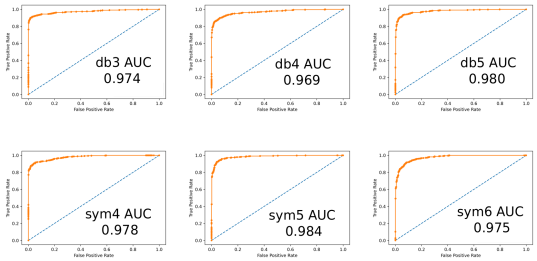


Figure 10: CNN based AUC-ROC curves for 2 class bird sound classification with wavelet based features

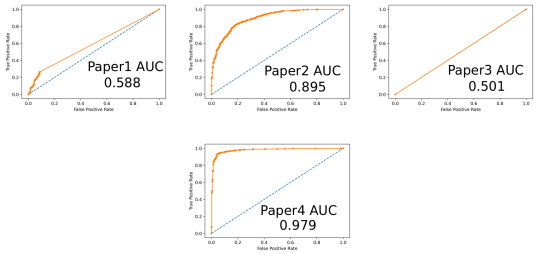


Figure 11: CNN based AUC-ROC curves for 2 class bird sound classification with features used in previous works

separability with an AUC = 0.58, Paper3 shows almost no skill. Finally, considering the Fourier based methods, it is observed that MFCC performs admirably with a near perfect score of 1 while both MFBE(0.962) and SPEC(0.986) show a big

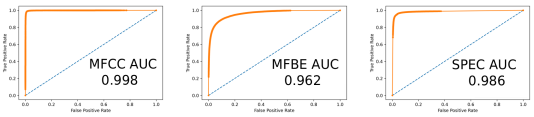


Figure 12: AUC-ROC curves for 2 class bird sound classification with fourier based features using CNN

improvement in their scores. Regarding the Fourier based methods, the computation of features such as MFCC and MFBE is motivated by human perception of sound. In order to compute MFCC, we made use of the DCT and this is due to the limitations of several ML algorithms. DCT is used to decorrelate the MFBE features (also called whitening). With the development of better ML algorithms and the growth of Deep Learning, the data is better learnt by the classification methods.

This insight can be **clearly seen with the results of MFCC and MFBE for both SVM and CNN.** While MFBE does not perform well with SVM as a classifier, when using CNN, the MFBE results are on par with MFCC! The same holds true for MFCC vs SPEC results too. Considering this, it is observed that unlike the Fourier based methods, the wavelet based methods are much simpler features that perform well over a wide variety of classifiers.

The next set of experiments are based on adding Additive White Gaussian Noise (AWGN) to a part of the audio dataset in order to understand how robust an audio feature is to noise. We consider four such levels, 0.2 AWGN (20% of the audio data is added with AWGN), 0.4 AWGN, 0.6 AWGN and 0.8 AWGN respectively. The results of these above experiments using the SVM classifier are seen in Fig.13. Given that the AWGN generated noise is completely random, when added to the original audio, it is understandable that the t-SNE plots does not depict any visible patterns for differentiating between the two classes. Beginning with the wavelet based features, namely daubechies (3,4) and symlets (5), the three features perform with an ROC of 0.96-0.97 with the 0 AWGN (original signals without any noise added). With 0.2 AWGN added (i.e. 20 % of the dataset is added with AWGN) to the dataset, the performance is still good across all the daubechies features with an ROC score of 0.9. As we increase the AWGN to 0.4, the performance almost remains constant with only a slight drop to about ROC =0.85. At 0.6 and 0.8 AWGN, there is a sharper relative degradation in the performance of the daubechies features than before as they score between ROC 0.64-0.66. One explanation for this good performance could be that since wavelets localize features in the audio data to different scales, the wavelets can preserve crucial audio features while removing noise.

Considering the fourier based methods, at 0 AWGN, there is a large margin between MFCC and the filter bank features (MFBE) and SPEC. While MFCC performed the best with the initial classification task scoring ROC = 1, the other two features scored in the range of 0.75-0.76. With an increase in noise added, at 0.2 AWGN, there is a sharper drop with the MFCC features (10% absolute reduction) when compared to the other wavelet and fourier features. However, apart from MFCC, considering MFBE and SPEC, it is observed that the drop in performance from 0 to 0.2 AWGN in the case of MFBE is negligible while in the case of SPEC, there is a drop from 0.75 to 0.68. With an increase to 0.4 AWGN, the MFCC still performs consistently with an ROC =0.87, MFBE drops slightly from 0.75 to 0.73 while SPEC drops from 0.68 to 0.64. As the AWGN ratio is increased to 0.6, the performance of all three Fourier-based methods dropped by 0.1 while at 0.8 AWGN, MFCC was seen to degrade the most relatively from 0.6 AWGN (although still performing better compared to the other two). Considering the reference works, the performance of the features is a mixed bag. While none of the features perform as well as MFCC or the wavelet based features, they remain in the middle ground. While p2 and p4 perform well with an increase in AWGN, it is seen that p3 shows a lot of degradation with an increase in noise. This shows that it is not a suitable method for classification although it performs well with clean audio. The results obtained for p3 in CNN add substance to this argument. p1 too, performs at AUC 0.84 with clean audio data. With an increase in AWGN, it is observed that its performance is worse than SPEC based features - which shows much better performance beginning at a lower value at 0 AWGN (0.84 to 0.52 for p1 vs 0.75 to 0.53 for SPEC). The poor performance of p3 is again seen in the CNN results with no separability.

There are three observations/hypothesis that are made from these results.

- **Does not work on SVM but on CNN** : Better ML algorithms lead to better understanding of data by the classification methods than seen previously (as indicated by MFBE performing poorly on SVM but performing on par with MFCC on CNN).
- **Simpler method that works for all** : Unlike the Fourier methods, the wavelet based methods are simpler features (since they work for SVM too) that perform well over a wide variety of classifiers.
- **Does not work on CNN but on SVM** : Can we hypothesize that a good performance on the CNN classifier could mean better robustness to noise (and vice versa as seen in paper1

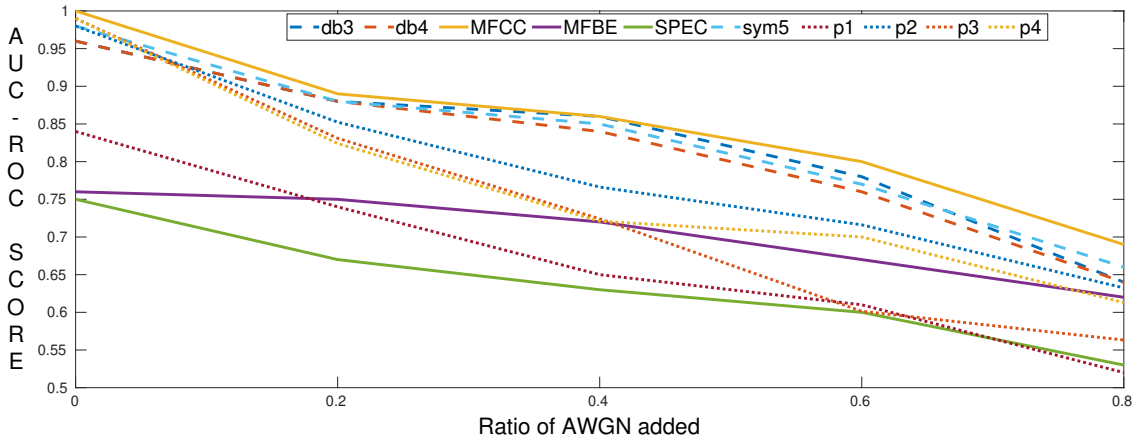


Figure 13: Performance of audio features at different AWGN ratios. The X axis depicts the ratio of AWGN added (0 to 80%) and the Y axis depicts the AUC-ROC score.

and paper3)? This could be studied in detail in the future.

X. Conclusions

This work being based on understanding Wavelets and Sparse Signal Representations, an attempt has been made into comparing and understanding how wavelet based features such as daubechies and symlets perform against fourier based methods in audio processing. The Receiver Operating Characteristic curve (ROC) has been employed to test the diagnostic ability of the audio features. We began with mainly two questions.

- Can wavelet based features hold their own against the current state of the art methods?
- Can wavelet features contain good discriminating information?

The answer to both is a resounding yes. Along with the success of using wavelet based features for bird sound classification, **we have also been able to improve upon the benchmark AUC-ROC score of 0.882** [12]. Wavelet features such as **db5 (0.97)** and **sym6 (0.99)** performed on a level similar to the best performing Fourier based methods (MFCC: 1.0) irrespective of the classifier used highlighting the possibility of using such wavelet based features for future audio processing experiments.

The project data can be accessed here

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] M. Bitenc, D. Kieffer, and K. Khoshelham, "Evaluation of wavelet denoising methods for small-scale joint roughness estimation using terrestrial laser scanning.," *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 2, 2015.

[14] R. Ramos, B. Valdez-Salas, R. Zlatev, M. Schorr Wiener, and J. M. Bastidas Rull, "The discrete wavelet transform and its application for noise removal in localized corrosion measurements," *International Journal of Corrosion*, vol. 2017, 2017.

[15] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE transactions on image processing*, vol. 9, no. 9, pp. 1532–1546, 2000.

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

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

[18] F. Chollet *et al.*, "Keras." https://keras.io, 2015.

[19] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.