Augmenting Audio-based Bird Species Identification with Music Processing Approaches

Aneesh Vartakavi Alexander Lerch
Georgia Tech Center for Music Technology
Georgia Institute Of Technology
Atlanta, Georgia(U.S.A)

13th Jul, 2014

Abstract

The task of automatic identification of bird species has steadily gained interest over the last few years, correlating with advances in allied fields such as Machine Learning and Pattern Recognition. A tool designed for this task could assist in identification and conservation efforts, reducing human involvement and effort in the process. Most prior work approaches the problem by parameterization of bird vocalizations, which has traditionally been restricted to low level features like the Spectral Centroid or MFCCs extracted from audio. Motivated by the perceptual and musical attributes of bird vocalizations, this study attempts to explore a large number of features in the dimensions of rhythm, timbre and pitch, which, to the best of the authors' knowledge, is the first study of this scale. An effective unsupervised denoising process is presented as a pre-processing step inspired by music signal processing. This work also evaluates the performance of feature categories and multiple supervised classification algorithms on the BirdCLEF dataset - a publicly available dataset of field recordings containing 501 species of birds, using commonly used performance metrics in the field of Machine Learning.

Keywords. Automatic Bird Species Identification, Machine Learning, Audio Content Analysis

1 Introduction

A computational system for the automatic identification of bird species could assist ecological conservation and monitoring efforts with remote beacons since birds are important markers of biodiversity (Briggs, Lakshminarayanan, Neal, Fern, Raich, Hadley, Hadley, and Betts 2012). A system robust to noise could help identify birds in field conditions, which could be useful when visual identification is not feasible or when different species look identical, while being non-intrusive. It could also be a valuable recreational and learning tool, a report by the US Fish and Wildlife Service estimates that about 48 million people in the USA identify as 'birders', with the industry generating 82 billion dollars of output (Fish and Service 2006).

The task of automatic identification of bird species through audio is related to the field of Musical Information Retrieval (MIR), concerned with the extraction of information from audio or symbolic data. The field frequently deals with the analysis of musical attributes like pitch, rhythm and timbre, and many algorithms have been developed with varying degrees of success. Some musicologists claim that bird song had a large influence in the development of music (Head 1997), and many classical composers created works directly inspired by bird song - *The Goldfinch*, a flute concerto by Vivaldi, *The Cuckoo and the Nightingale*, an organ concerto by Handel, and *Catalogue d'Oiseaux* by Messiaen are examples of the latter. Humans and birds themselves identify bird species through musical attributes such as frequency range, trills, whistles and others (Sinnott 1980). This connection suggests that tools and approaches in music analysis might benefit the task of bird species identification. We therefore investigate "musical features" derived from MIR work such as musical style recognition and how they can be used to improve current bird classification systems, potentially the first study of this nature.

The following section puts this work in context with prior work in the field. Section 3 details the system that was developed for this task, while Section 4 presents and discusses the results obtained.

2 Related Work

The automatic identification of bird species has drawn increasing interest over the years. In this section, we highlight and discuss some previous approaches. The common processing steps of audio-based bird species classification are

- *pre-processing*: processing the input audio to enhance quality and to unitize the input audio format,
- *segmentation*: isolating syllables of bird calls in the recording,
- *feature extraction*: extracting descriptors which represent the audio content, and
- *classification*: the automatic bird categorization.

The necessary pre-processing steps vary largely in scope and are dependent on the input signal properties and the chosen data set. Some early approaches used audio recorded in laboratory conditions (Anderson, Dave, and Margoliash 1996; Kogan and Margoliash 1998), while others used field recordings. Briggs et al., process field recordings by discarding frequencies below 1.378 kHz and above 10.852 kHz in their spectrogram, and use magnitude based thresholding to determine 'interesting frames' (Briggs, Raich, and Fern 2009). Fodor performs Gaussian filtering and morphological opening and closing on the spectrogram for noise removal (Fodor 2013). Wavelet denosing by hard thresholding was used by Sun and Mita (Sun and Mita 2012). Unrelated to bird species identification, Vaseghi (Vaseghi 2008) and Boll (Boll 1979) discuss spectral subtraction approaches to denoise audio, which was used in the preprocessing stage of this work.

A few systems rely on a manual segmentation stage (Lopes, Lameiras Koerich, Nascimento Silla, and Alves Kaestner 2011; Selouani, Kardouchi, Hervet, and Roy 2005). Since a non-automatic segmentation is obviously impractical for large scale experiments, some systems attempt to segment the audio file automatically. Chen and Maher describe an approach using spectral peak tracking in which they track peaks from one spectral frame to the next (Chen and Maher 2006). Fagerlund proposes an interactive thresholding algorithm in the time domain (Fagerlund 2004). Briggs et al. note that energy based time domain segmentation degrades in noisy conditions and performs poorly when multiple bird calls overlap in time; they use a supervised segmentation using a Random Forest classifier, depending on annotated spectrograms for training (Briggs, Lakshminarayanan, Neal, Fern, Raich, Hadley, Hadley, and Betts 2012).

Mel Frequency Cepstral Coefficients (MFCCs) — a standard feature set in speech and music processing (Logan 2000) — have been successfully applied to the task by many research groups (Kogan and Margoliash 1998; Lee, Lee, and Huang 2006; Chen and Maher 2006; **?**; Graciarena, Delplanche, Shriberg, Stolcke, and Ferrer 2010; Fagerlund 2007). Lee et al. extract features from static and dynamic two-dimensional MFCCs (Lee, Han, and Chuang 2008). Lopes et al. (Lopes, Lameiras Koerich, Nascimento Silla, and Alves Kaestner 2011) evaluate the performance of feature sets extracted by existing tools, including Marsyas (Tzanetakis and Cook 2000), Sound Ruler (Bee 2004), and an Inter Onset Interval (IOI) histogram coefficient feature set as introduced by Gouyon et al. (Gouyon, Dixon, Pampalk, and Widmer 2004). Briggs et al. extract features such as area, perimeter and bandwidth from the segments at the output of their segmentation algorithm (Briggs, Lakshminarayanan, Neal, Fern, Raich, Hadley, Hadley, and Betts 2012). Tzanetakis et al. describe extracting timbral, rhythmic and pitch features for musical genre classification (Tzanetakis and Cook 2002).

Some early classification approaches include using Hidden Markov Models and Dynamic Time Warping to distinguish between two species of birds with distinct calls recorded in a low noise environment (Anderson, Dave, and Margoliash 1996; Kogan and Margoliash 1998). A few systems reported success Artificial Neural Networks (Selouani, Kardouchi, Hervet, and Roy 2005; McIlraith and Card 1997). Garciareana et al. built a front end optimization system using a Gaussian Mixture Model (Graciarena, Delplanche, Shriberg, Stolcke, and Ferrer 2010), also used in other systems (Kwan, Ho, Mei, Li, Ren, Xu, Zhang, Lao, Stevenson, Stanford, et al. 2006; Tyagi, Hegde, Murthy, and Prabhakar 2006). Fagerlund et al. used a Support Vector Machine(SVM) based system, creating a decision tree with SVM classifiers at each node. (Fagerlund 2007). Briggs et al. introduce and evaluate Multi-Instance/Multi-Label (MIML) classifiers; an MIML framework represents objects to be classified as a 'bag-of-instances' associated with multiple class labels (Briggs, Lakshminarayanan, Neal, Fern, Raich, Hadley, Hadley, and Betts 2012). In contrast, a Single-Instance/Single-Label classifier assigns exactly one class label to each observation. They conclude that MIML classifiers have better performance in their evaluation setup, and noting that annotation for a MIML classifier is less intensive, requiring only a list of species present in a recording and not a detailed annotation of each segment.

It is difficult to compare the results of the approaches detailed above, as they use different datasets with largely varying numbers of species and training examples. Lopes et al. compile accuracy results from some prior approaches, also making a similar observation (Lopes, Gioppo, Higushi, Kaestner, Silla, and Koerich 2011). Prior approaches also have fundamental differences - for example Towsey et al. (Towsey, Planitz, Nantes, Wimmer, and Roe 2012) use a bird-call dependent feature set and a simple classifier, while most other works surveyed use a constant feature set and a complex classifier. A recent surge in data prediction competitions based on this task allow for comparison of results by providing a sandbox to test and benchmark multiple algorithms. The MLSP-2013[1] and NIPS-2013[2] competitions focused on multi-label approaches with 19 and 87 classes of birds, respectively. ICML-2013[3] included a bird species identification task with 36 classes of birds, and ICML-2014 extends the previous version by posing it as an unsupervised learning problem. The BirdCLEF-2014[4] task utilizes data from Xeno-Canto[5], a popular online repository of user submitted recordings of bird species. The subset released for this task was used to evaluate the system, comprising of 501 species of birds and a total of about 14,000 recordings.

3 System Description

The structure of our system mirrors a system with the processing steps above, which is also similar to the structure of many MIR systems. The current system is designed to be a Single-Instance/Single-Label approach, and the main species of the audio file is assumed to be the only species present in the recording.

3.1 Pre-Processing

Since we have to deal with field recordings we can expect varying audio quality due to atmospheric conditions like wind and rain, interfering bird and insects calls, varying quality of the recording equipment, and varying professionalism of the recordist. An effective denoising procedure is therefore necessary. The input audio is downsampled to 16 kHz, and the spectrogram is computed with a block size of 1024 and 50% overlap. A two stage denoising process based on spectral subtraction is applied to the audio signal with two different noise estimates. The first stage removes the median of each frequency bin across time frames. For the second stage, the noise is estimated using the kMeans algorithm. More specifically, the centroid of the largest cluster of three clusters of spectral magnitudes is assumed to be the noise estimate for spectral subtraction. Thresholding is then performed at -40dB below the maximum, completing the pre-processing stage.

The panels of the Fig. 1 (top down) display the original spectrogram, the denoised spectrogram after each of the pre-processing stages, and the spectrogram after the segmentation stage (Sect. 3.2). A bounding rectangle is drawn around each connected component in the final panel to illustrate the output of the segmentation process, described below.

3.2 Segmentation

Detecting connected components is a method frequently used in image processing to identify and group pixels according to their position and values. Two pixels are assumed to be connected if they satisfy the properties of 8-connectivity (see Fig. 2).

We detect connected components in the denoised spectrogram and discard all components for which the values for area and weighted centroid fall below empirical thresholds. This

[1] `www.kaggle.com/c/mlsp-2013-birds`, accessed on 13th Jul, 2014

[2] `www.kaggle.com/c/multilabel-bird-species-classification-nips2013`, accessed on 13th Jul, 2014

[3] `www.kaggle.com/c/the-icml-2013-bird-challenge`, accessed on 13th Jul, 2014

[4] `www.imageclef.org/2014/lifeclef/bird`, accessed on 13th Jul, 2014

[5] `www.xeno-canto.org`, accessed on 13th Jul, 2014

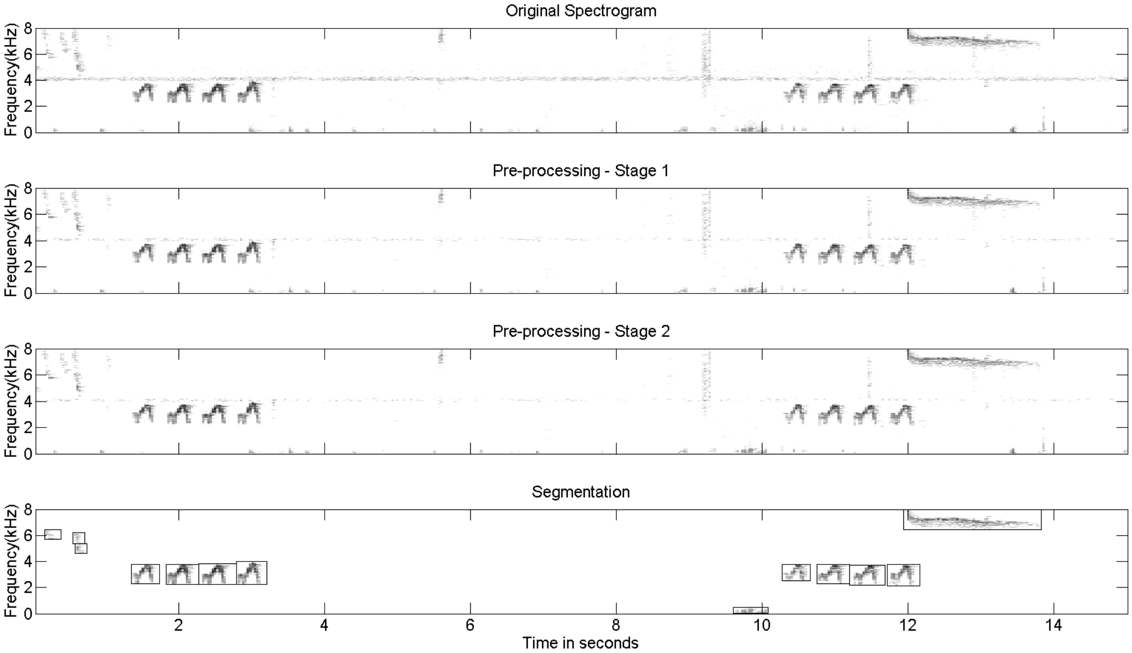

Figure 1: Output spectrograms of individual stages in the pre-processing algorithm

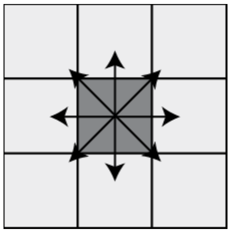

Figure 2: 8-Connected Neighborhood - The pixels marked by arrows are connected to the pixel in the center.

technique is similar to 'Acoustic Event Detection' as described by Towsey et al. (Towsey, Planitz, Nantes, Wimmer, and Roe 2012).

3.3 Feature Extraction

Following pre-processing, features are extracted on the sample. The investigated feature set can be roughly divided into five major groups.

3.3.1 Connected Component Features

For each connected component (see Sect. 3.2), the following features are extracted: In this stage, features from the connected components extracted in the pre-processing stage are extracted; see Briggs et al. for the motivation for some of the features (Briggs, Lakshminarayanan, Neal, Fern, Raich, Hadley, Hadley, and Betts 2012).

- *Area*: number of pixels
- *Perimeter*: number of pixels on the component border
- *Weighted Centroid* - mass center of the component based on location and intensity of pixels
- *Time Duration*: number of time frames spanned by the component
- *Frequency Bandwidth*: number of frequency bins spanned by the component
- *Ratio of lengths of the major and minor axis* : Properties of an ellipse with the same normalized second moments as the component
- *Orientation*: angle between the abscissa and the major axis
- *Non-Compactness*: squared perimeter length divided by the area (Briggs, Lakshminarayanan, Neal, Fern, Raich, Hadley, Hadley, and Betts 2012)
- *Rectangularity*: area divided by bandwidth times duration (Briggs, Lakshminarayanan, Neal, Fern, Raich, Hadley, Hadley, and Betts 2012)
- *Intensity Descriptors*: mean, maximum, and minimum magnitude plus the standard deviation of the magnitude

The mean and standard deviation of all these features for all connected components of one recording are used to represent the audio file with a 26-dimensional feature vector.

3.3.2 Timbre Features

This feature category consists of instantaneous spectral features computed per time frame, mostly representing the timbral properties of the signal. The features in this set calculated per audio frame and are traditionally the most commonly used low level features in many audio classification systems. Detailed definitions for the spectral features can be found in (Lerch 2012).

- *Max Frequency*: maximum frequency present in the time frame
- *Bandwidth* The difference between maximum and minimum frequencies

- *Spectral Centroid*: the centroid of the magnitude spectrum
- *Spectral Crest*: a measure of sinusoidality
- *Spectral Decrease*: a measure of spectral envelope shape
- *Spectral Flatness*: a measure of noisiness
- *Spectral Flux*: a measure of spectral change
- *Spectral Kurtosis*: a measure of Gaussianity
- *Spectral Rolloff*: a measure of bandwidth
- *Spectral Skewness*: a measure of skewness
- *Spectral Slope*: a measure of spectral shape
- *Spectral Spread*: a measure of instantaneous bandwidth
- *16 MFCCs*: mel frequency cepstral coefficients according to the specifications provided by BirdCLEF

Feature values for the silence frames are discarded in order to reduce bias. The mean and standard deviation of all these features and their derivatives (in the case of the MFCCs: also the 2nd derivative) across time frames are used to represent the audio file. The number of MFCC features is 96 and the number of the remaining features is 48.

3.3.3 Onset Features

An onset marks the begin of a musical note or event, and a pattern of onsets may create rhythm. 'Onset features' are therefore used to describe the rhythm space of an audio signal. We use a standard method for onset extraction based on spectral flux (Bello, Daudet, Abdallah, Duxbury, Davies, and Sandler 2005), and compute the histogram of Inter Onset Intervals (IOIs) for each recording (compare (Gouyon, Dixon, Pampalk, and Widmer 2004)). The IOI histogram is created with 16 bins of width 125 ms each. The following features are computed from the IOI histogram:

- *Standard Deviation*
- *Ratio of first and second maximum (magnitude and indices)*
- *Number of Zero bins*
- *Centroid*
- *Crest*
- *Decrease*
- *Kurtosis*
- *Rolloff*
- *Skewness*
- *Slope*
- *Spread*

This results in a 12-dimensional Onset-feature vector per file.

3.3.4 Pitch Features

Pitch is the subjective perception of frequency - the fundamental frequency and its contour are represented in this feature category. We use Yin (De Cheveigné and Kawahara 2002) for frequency estimation, a popular pitch estimation algorithm, computed with default parameterization. Results outside connected components and frequency results outside a range of 100 Hz–8 kHz are discarded. For each segment, a third order polynomial is fitted to the derivative of the frequency contour. The mean and standard deviation of the resulting four coefficients across all segments are stored, along with the values of the mean and standard deviation of the frequency contour itself across time leads to a 10-dimensional pitch feature vector for each file.

3.3.5 Meta Data Features

The BirdCLEF dataset provides some meta data with each recording, which we included and described as 'meta data features' in our set:

- *Hour of recording*
- *Month of recording*
- *Latitude and Longitude*
- *Elevation*

3.4 Classification

We use Weka (Hall, Frank, Holmes, Pfahringer, Reutemann, and Witten 2009), a popular machine learning and data mining software for this stage. Weka provides convenient access to many different types of classification algorithms, as well as an interface to visualize data.

3.4.1 Classifiers

We investigate the following classifiers:

- Zero-R: baseline classifier that simply always predicts the majority class
- One-R: single level decision tree with one rule. This classifier is often surprisingly accurate and can be easily interpreted by humans (Buddhinath and Derry 2006).
- Nearest Neighbor: remembers all training data points with their classes uses the closest class label as estimate
- Linear SVM: Support Vector Machine with a linear kernel. We used Weka's wrapper for LibLINEAR, a library for large linear classification (Fan, Chang, Hsieh, Wang, and Lin 2008); we use both terms interchangeably.
- Random Forest: ensembles of decision trees.

4 Results and Discussion

The following section discusses results obtained using the BirdCLEF dataset, all results are obtained using 10-fold cross validation.

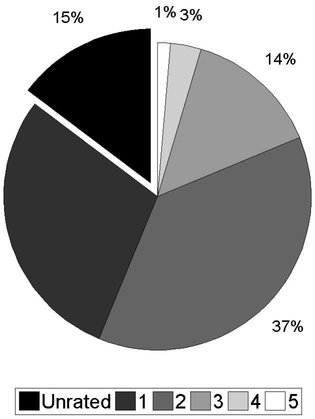

Figure 3: Distribution of user submitted quality ratings

4.1 The BirdCLEF Dataset

The BirdCLEF dataset is the first attempt to create a large, public dataset for the task of automatic bird species identification. The dataset consists of 501 species of birds from South America (some recordings were labeled *unknown species*). The training set consists of 9688 recordings and meta data associated with each recording, along with a set of MFCCs and their first two derivatives.

The dataset is sourced from the Xeno-Canto database, a large online collection of non-moderated user-submitted recordings, and therefore suffers from inconsistent audio quality. The recordings are also subject to varying atmospheric conditions like rain or wind noise, or handling noise. Furthermore, many instances in the database contain background bird species, and many are tagged as such. However, the absence of this information could mean both that the recording does not have any background species, or that the recordist is unsure of the species or did not consider them important.

Xeno-Canto allows the recordist to upload comments, some comments contain valuable information about the age of the birds, the number of birds or the type of call. The lack of a standard format, however, makes extracting relevant information automatically on a large scale difficult. For example, the database lacks a standard format for time, containing entries like '?', '8am', '11:00' and 'morning'.

The recordists can also rate their recordings in the process of uploading to the Xeno-Canto database, but there are no clear guidelines on how to score recordings, leaving the resulting ratings prone to subjective bias. A pie chart of the quality scores is presented in Figure 3, with one being the best quality and 5 the least. It can be seen that the number of unrated recordings is high, and that the distribution is asymmetrical, with a lot more recordings with high ratings and a large drop between the number of recordings rated 2 and 3.

The dataset averages about 19.33 recordings per species, with a standard deviation of 8.19. The maximum number of instances per species is 62 while the minimum is 11.

4.2 Evaluation Metrics

In order to make the results presented below comparable to both previous research in the field and to evaluation metrics commonly used in competitions, we use a number of metrics. All metrics except the micro and macro averaged F-Measure, are averaged and weighted by the number of instances in each class, as calculated by Weka. Some metrics are mathematically formulated using True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). The metrics used are:

- Percentage Accuracy: (TP + TN) / (TP + TN + FP + FN)
- Precision: (TP) / (TP + FP)
- Recall: (TP) / (TP + FN)
- Area Under ROC curve (AUC): The ROC curve is a convenient method to visualize the performance of classifiers, the AUC condenses the information into a single figure of merit (Bradley 1997).
- Area under Precision Recall Curve (PRC): The area under the PRC provides a more informative picture of an algorithm's performance on a skewed dataset (Davis and Goadrich 2006).
- Unweighted Micro Averaged F-Measure (Micro-F): The micro averaged F measure is biased towards the classifier's performance on common categories.
- Unweighted Macro Averaged F-Measure (Macro-F): The macro averaged F measure is dominated by the classifier performance on rare categories (Yang and Liu 1999).

The percentage accuracy is arguably the most intuitive of the performance metrics, but could incorrectly represent classifier performance in skewed datasets; precision and recall are more useful measures in this context. The AUC has widely been adopted as a more statistically consistent and discriminative measure than accuracy (Huang and Ling 2005), although some claim it is flawed (Lobo, Jiménez-Valverde, and Real 2008) and caution against it's use.

4.3 Pre-Processing Evaluation

A qualitative visual examination of the pre-processing and segmentation stage verifies that the process removes most of the noise as shown in Fig. 1.

In order to quantify the improvement in classification accuracy with and without pre-processing and segmentation, the MFCC features are computed with and without denoising and segmentation. Table 1 displays the results for the percentage accuracy (standard deviation in parentheses) using the 30 species subset. The Random Forest was built with 200 trees, constructed while considering 10 random features. It can be observed that classification accuracy after pre-processing increases accuracy by a large margin regardless of the classifier. These results emphasize the importance of a powerful pre-processing stage for good classification results.

Classifier	No Pre-processing	With Pre-processing
ZeroR	6.81 (0.06)	6.81 (0.06)
kNN (k=1)	25.21 (5.33)	48.74 (6.60)
LibLINEAR	43.64 (5.99)	61.67 (5.32)
Random Forest	34.42 (5.55)	57.06 (5.53)

Table 1: Percentage accuracy of a 30 species subset with and without pre-processing for different classifiers

Classifier	Accuracy (%)	Precision	Recall	AUC	PRC	Macro-F	Micro-F
ZeroR	0.64 (0.04)	0.0	0.01	0.5	0.0	0.0	0.01
OneR	1.9 (0.46)	0.02	0.02	0.51	0.0	0.0	0.02
kNN(k=1)	17.39 (1.01)	0.17	0.17	0.59	0.10	0.14	0.17
SVM	31.10 (1.59)	0.28	0.31	0.65	0.17	0.24	0.31
RF	24.61 (1.07)	0.20	0.25	0.89	0.3	0.17	0.24

Table 2: Evaluation results for the full data set for selected classifiers

4.4 Full System Evaluation

The results for the complete system, using all pre-processing stages and the complete feature set and using the complete training set with 501 classes are are summarized in Table 2. The Random Forest was built with 200 trees considering 15 random features, and was restricted to a depth of 10.

It can be observed that the Linear SVM excels in classification accuracy, but the Random Forest has a higher AUC and larger Area Under PRC. The Random Forest is therefore a more discriminant classifier; it is able to identify a random positive sample with greater probability than a random negative sample (Bradley 1997).

Briggs et al. report a micro-AUC of 0.949 achieved with an SISL-Random Forest on 13 species of birds (Briggs, Lakshminarayanan, Neal, Fern, Raich, Hadley, Hadley, and Betts 2012). We chose a subset of 13 species in alphabetical order from the BirdCLEF dataset, and applied a Random Forest considering 12 random features and 150 trees. We achieved a macro averaged AUC of 0.94 - however, the number of training instances per species and the species itself could effect results.

4.5 Feature Categories

To shed some light on the contribution of the individual feature categories on the overall result, the percentage accuracy is computed for each category. The results for the 30 species subset of the data are shown in Fig. 4. The Random Forest was built with 150 trees, allowing Weka to choose the number of random features. For this plot, the timbral features category has been split into MFCC features and remaining spectral features to allow for comparisons

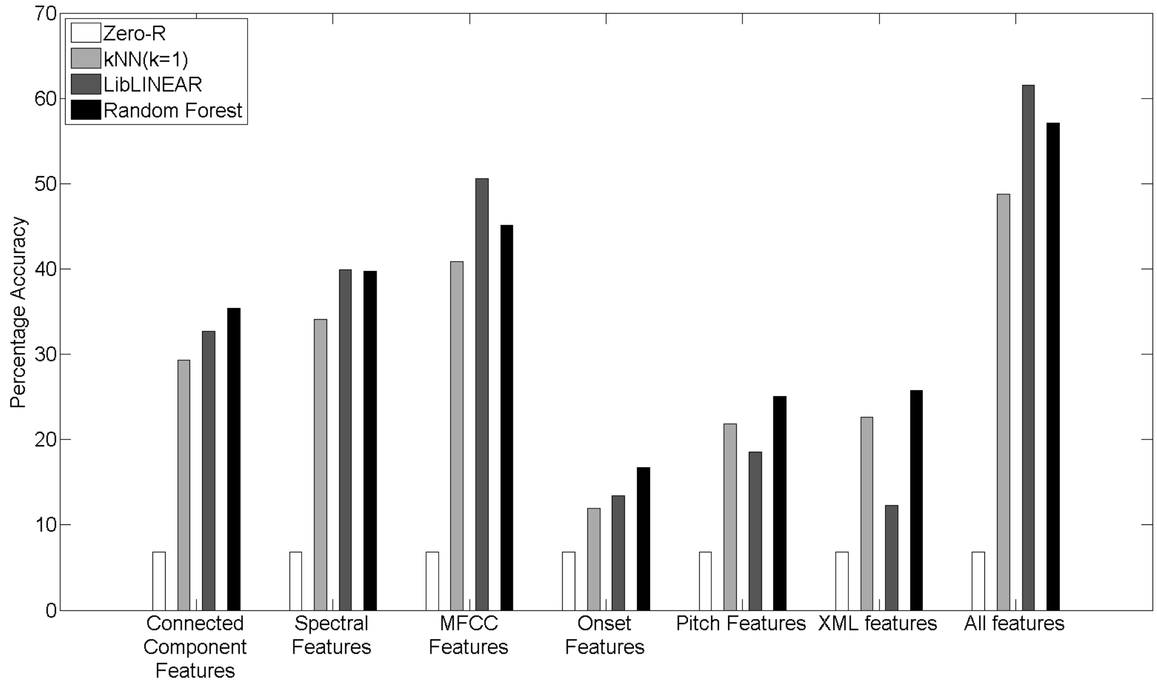

Figure 4: Evaluating Feature Categories for a subset of thirty features

with other approaches using only MFCC features.

It can be observed that the Random Forest performs comparable to or better than the SVM, with the exception of the full feature set and the MFCC features, where the SVM outperforms the Random Forest classifier. The MFCC features have the highest accuracy within the feature categories. The pitch features and especially the rhythm features perform significantly worse than the other feature categories (except the XML features with their expected poor performance). There are three possible explanations for this

- *dataset*: The dataset contains multiple types of bird calls and songs, resulting in drastically different file lengths. Figure 5 displays a box-plot of the file lengths for a subset of fifty species chosen in alphabetical order, clipped to a length of 120 s for better visualization. The standard deviation can be seen to be comparable in value to the mean of the file lengths; there are very short as well as long recordings. The poor performance of the rhythm features could potentially be attributed to insufficient parametrization of the rhythm (and pitch) space.

- *feature extraction*: If the onset detection method for unknown reasons gives erratic results, both the rhythm features as well as the dependent pitch features will show poor performance.

- *bird vocalization structure*: The variety of different syllables and phrases (and pauses) in bird calls might make it difficult to represent one class with all the features - our parameterization of the rhythm and pitch feature space is probably insufficient. Timbre features, on the other hand, are per definition robust against such variations in pitch

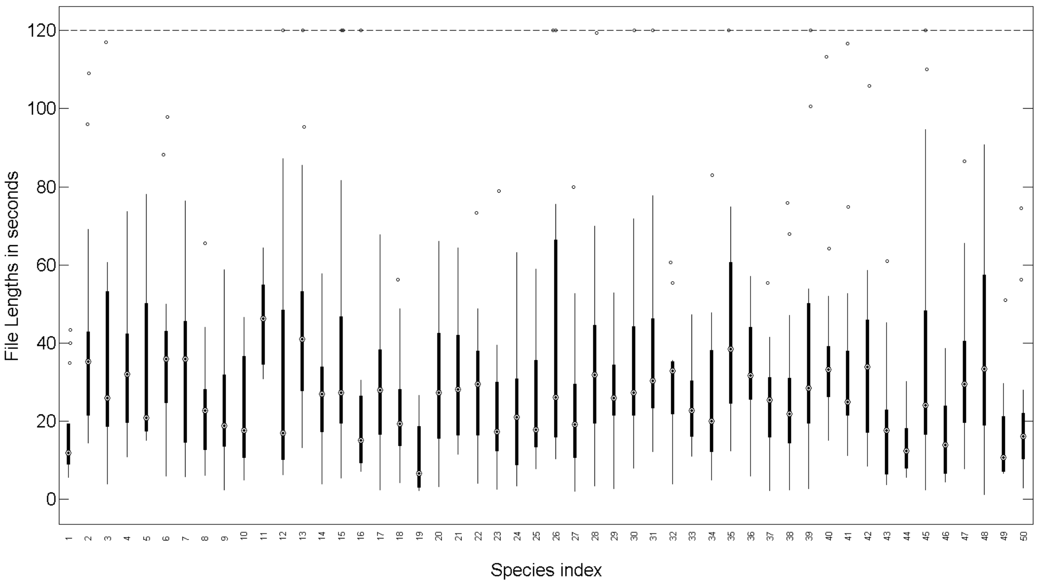

Figure 5: Box plot of the file lengths of 50 species from the BirdCLEF dataset

and rhythm of the bird call.

4.6 Varying Test Set Size

The number of classes(species of birds) the classifier is trained with impacts its performance reflected in metrics such as the percentage accuracy. We created subsets by choosing the required number of species in alphabetical order. The comparison is visualized in Fig. 6, with the elements representing percentage accuracy. The Random Forest was built with 200 trees, considering 15 random features and the depth restricted to 10.

As expected, performance decreases rapidly with an increase in the number of classes. The linear SVM has the highest accuracy in all cases, with the Random Forest and the kNN following.

4.7 Dimensionality Reduction

In order to check for redundancy in the feature set, Principal Component Analysis was applied. We found 40 principal components with eigenvalues greater than 1. All the feature groups were clearly represented in these principal components, with timbre features dominating the majority.

5 Conclusions

In summary, this work applied music-inspired pre-processing and introduced high level musical features such as pitch and rhythm descriptors to aid the automatic classification of bird species through audio. The unsupervised denoising scheme was verified to be effective. The BirdCLEF dataset, containing 501 species of birds was used, with Linear SVM's proving to have the

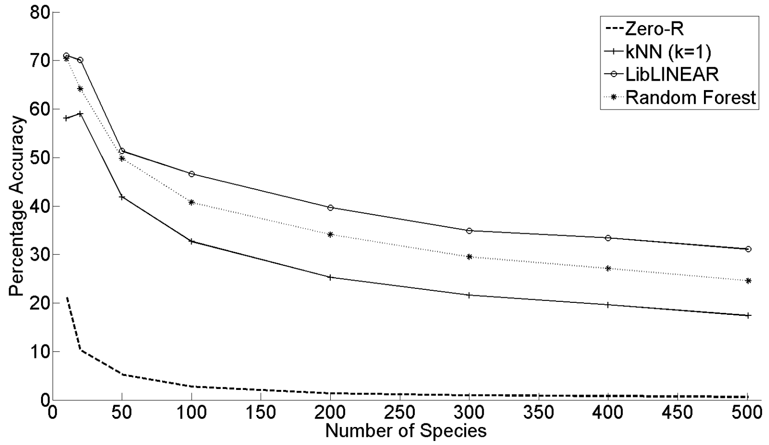

Figure 6: Percentage accuracy with varying subset sizes

highest classification accuracy among the classifiers tested. Of the various categories of features tested, the MFCC features had the highest performance within the feature categories. The combination of all the features representing the timbre, pitch and rhythm dimensions achieved the highest classification accuracy.

The task of automatic identification of bird species through audio is garnering research interest in the last few years. The task is inherently challenging because of the large number of vocalizations of a single species, coupled with factors like the regional variations of bird song with geographical region, the correlations between closely related species, seasonal variations in bird song and temporal plasticity. Yet, systems are achieving increasing levels of success, and great progress can be expected in the next few years.

While current systems use sophisticated classifiers, we believe there is potential for future work in the design of new features to describe the pitch and rhythm dimensions and in improving the pre-processing stage. A hierarchical classifier could potentially make use of information from the phylogenetic tree, and discriminate between bird species more effectively by building carefully parametrized classifiers at each node. This might especially be important in the case of large numbers of classes and closely related bird species. Attribute selection and dimensionality reduction were briefly only explored in this work, but a deeper look could help remove redundancy in the feature set. If the resources are available, a supervised pre-processing scheme could also help improve performance. A different approach is explored by the ICML-2014 competition by requiring unsupervised classification. Finally, Multi-Instance/Multi-Label classification approaches are worth exploring since they reflect the nature of practical use cases of automatic bird classification systems well.

References

Anderson, S. E., A. S. Dave, and D. Margoliash (1996). Template-based automatic recognition of birdsong syllables from continuous recordings. *The Journal of the Acoustical Society of America 100*(2), 1209–1219.

Bee, M. (2004). Sound ruler acoustical analysis: A free, open code, multi-platform sound analysis and graphic package. *Bioacoustics 14*, 171–178.

Bello, J. P., L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler (2005). A tutorial on onset detection in music signals. *Speech and Audio Processing, IEEE Transactions on 13*(5), 1035–1047.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on 27*(2), 113–120.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition 30*(7), 1145–1159.

Briggs, F., B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. G. Betts (2012). Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America 131*(6), 4640–4650.

Briggs, F., R. Raich, and X. Z. Fern (2009). Audio classification of bird species: A statistical manifold approach. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pp. 51–60. IEEE.

Buddhinath, G. and D. Derry (2006). A simple enhancement to one rule classification. *Technique Report at*.

Chen, Z. and R. C. Maher (2006). Semi-automatic classification of bird vocalizations using spectral peak tracks. *The Journal of the Acoustical Society of America 120*(5), 2974–2984.

Davis, J. and M. Goadrich (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240. ACM.

De Cheveigné, A. and H. Kawahara (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America 111*(4), 1917–1930.

Fagerlund, S. (2004). *Automatic recognition of bird species by their sounds*. Ph. D. thesis, Helsinki University of technology.

Fagerlund, S. (2007). Bird species recognition using support vector machines. *EURASIP Journal on Advances in Signal Processing 2007.*

Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research 9*, 1871–1874.

Fish, U. and W. Service (2006, March). Birding in the united states : A demographic and economic analysis.

Fodor, G. (2013). The ninth annual mlsp competition: First place. In *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pp. 1–2. IEEE.

Gouyon, F., S. Dixon, E. Pampalk, and G. Widmer (2004). Evaluating rhythmic descriptors for musical genre classification. In *Proceedings of the AES 25th International Conference*, pp. 196–204. Citeseer.

Graciarena, M., M. Delplanche, E. Shriberg, A. Stolcke, and L. Ferrer (2010). Acoustic front-end optimization for bird species recognition. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 293–296. IEEE.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter 11*(1), 10–18.

Head, M. (1997). Birdsong and the origins of music. *Journal of the Royal Musical Association 122*(1), 1–23.

Huang, J. and C. X. Ling (2005). Using auc and accuracy in evaluating learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on 17*(3), 299–310.

Kogan, J. A. and D. Margoliash (1998). Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: A comparative study. *The Journal of the Acoustical Society of America 103*(4), 2185–2196.

Kwan, C., K. Ho, G. Mei, Y. Li, Z. Ren, R. Xu, Y. Zhang, D. Lao, M. Stevenson, V. Stanford, et al. (2006). An automated acoustic system to monitor and classify birds. *EURASIP Journal on Advances in Signal Processing 2006*.

Lee, C.-H., C.-C. Han, and C.-C. Chuang (2008). Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients. *Audio, Speech, and Language Processing, IEEE Transactions on 16*(8), 1541–1550.

Lee, C.-H., Y.-K. Lee, and R.-Z. Huang (2006). Automatic recognition of bird songs using cepstral coefficients. *Journal of Information Technology and Applications 1*(1), 17–23.

Lerch, A. (2012). *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons.

Lobo, J. M., A. Jiménez-Valverde, and R. Real (2008). Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography 17*(2), 145–151.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, Plymouth.

Lopes, M. T., L. L. Gioppo, T. T. Higushi, C. A. Kaestner, C. Silla, and A. L. Koerich (2011). Automatic bird species identification for large number of species. In *Multimedia (ISM), 2011 IEEE International Symposium on*, pp. 117–122. IEEE.

Lopes, M. T., A. Lameiras Koerich, C. Nascimento Silla, and C. Alves Kaestner (2011). Feature set comparison for automatic bird species identification. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pp. 965–970. IEEE.

McIlraith, A. L. and H. C. Card (1997). Birdsong recognition using backpropagation and multivariate statistics. *Signal Processing, IEEE Transactions on 45*(11), 2740–2748.

Selouani, S., M. Kardouchi, E. Hervet, and D. Roy (2005). Automatic birdsong recognition based on autoregressive time-delay neural networks. In *computational intelligence methods and applications, 2005 ICSC congress on*, pp. 6–pp. IEEE.

Sinnott, J. M. (1980). Species-specific coding in bird song. *The Journal of the Acoustical Society of America 68*(2), 494–497.

Somervuo, P., A. Harma, and S. Fagerlund (2006). Parametric representations of bird sounds for automatic species recognition. *Audio, Speech, and Language Processing, IEEE Transactions on 14*(6), 2252–2263.

Sun, R. and N. Mita (2012). Nocturnal wild bird species identification by sound information using wavelet. In *Wavelet Active Media Technology and Information Processing (lCW AMTIP), 2012 International Conference on Date of Conference*, pp. 17–19.

Towsey, M., B. Planitz, A. Nantes, J. Wimmer, and P. Roe (2012). A toolbox for animal call recognition. *Bioacoustics 21*(2), 107–125.

Tyagi, H., R. M. Hegde, H. A. Murthy, and A. Prabhakar (2006). Automatic identification of bird calls using spectral ensemble average voice prints. In *Proceedings of the Thirteenth European Signal Processing Conference.*

Tzanetakis, G. and P. Cook (2000). Marsyas: A framework for audio analysis. *Organised sound 4*(3), 169–175.

Tzanetakis, G. and P. Cook (2002). Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on 10*(5), 293–302.

Vaseghi, S. V. (2008). *Advanced digital signal processing and noise reduction.* John Wiley & Sons.

Yang, Y. and X. Liu (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 42–49. ACM.