Georg-August-Universität Göttingen

Machine Learning-Based Analysis of Bird Vocalizations

Dissertation

for award of the doctoral degree
in Mathematics and Natural Sciences
"Doctor rerum naturalium"
of the Georg-August-Universität Göttingen

within the doctoral program
PhD Programme in Computer Science (PCS)
of the Georg-August University School of Science (GAUSS)

Submitted by:

Burooj Ghani

from Srinagar

Göttingen, 2021

Thesis Advisory Committe

Prof. Dr. Florentin Wörgötter
Bernstein Center for Computational Neuroscience, Third Institute of Physics, Georg-August-Universität Göttingen, Germany

Prof. Dr. Sarah Hallerberg
Faculty for Engineering and Computer Science, Hamburg University of Applied Sciences, Germany

Dr. Minija Tamosiunaite
Bernstein Center for Computational Neuroscience, Third Institute of Physics, Georg-August-Universität Göttingen, Germany

Reviewers and Members of the Examination Board

Prof. Dr. Florentin Wörgötter
Bernstein Center for Computational Neuroscience, Third Institute of Physics, Georg-August-Universität Göttingen, Germany

Prof. Dr. Sarah Hallerberg
Faculty for Engineering and Computer Science, Hamburg University of Applied Sciences, Germany

Other Members of the Examination Board

Prof. Dr. Carsten Damm
Faculty of Mathematics and Computer Science, Institute of Computer Science, Georg-August-Universität Göttingen, Germany

Prof. Dr. Marcus Baum
Faculty of Mathematics and Computer Science, Institute of Computer Science, Georg-August-Universität Göttingen, Germany

Prof. Dr. Jörg Enderlein
Third Institute of Physics, Georg-August-Universität Göttingen, Germany

Prof. Dr. Alexander Gail
Sensorimotor Group, German Primate Center, Göttingen, Germany

Date of Oral Examination: 15.02.2022

To the "ordinary" of the world,
ones for whom the dice did not roll favourably.

Preface

This dissertation embraces efforts that I put in during the last few years as a PhD student, jointly at the University of Göttingen and Hamburg University of Applied Sciences. I have been working under the supervision of Prof. Sarah Hallerberg. In this time frame I was for the first time exposed to the challenges of working on problems that require you to embrace the attitude of "thinking outside the box" and owing to the struggles and obstacles I faced along the way I have to say I discovered the stoic of my life. My research is all about sounds and birds. It is about teaching machines how to perceive sounds that birds produce. Clearly, the way I look at birds and their beautiful, myriadly intricate songs has changed a bit since I started working on this topic.

There are lot of people that, directly or indirectly, have contributed to this work. Although it would be impossible to mention every single one of them by name, I will make an attempt to highlight a few names. For believing in me since the very first moment I am thankful to my supervisor Sarah, who has been supportive all along. Thank you for all the zillion discussions, advice, support and encouragement. For being always ever generous and supportive in terms of feedback and steering I thank Prof. Florentin Wörgötter. Also, I wish to thank Dr. Holger Klinck for great collaboration and for always promptly making time for our fruitful meetings; Prof. Timo Gerkmann for hosting me at the University of Hamburg and sparing time for the detailed and insightful discussions. Thank you, Mr. Bieck, for helping me deal with the most difficult times during this journey – without your counseling and humongsly effective advice I would not have succeeded in finishing this thesis. You taught me resilience. Since my first step into this journey, I will not ever be thankful enough to Theresa, who believed in me and lent her unabated support during one of the most challenging times.

On the flatmates-side – Vivi, Denis, Ailen, Sven, Sara and Liam – I am immensely thankful to you all, for providing me a home. PhD is a lonely journey and having a place where you can feel home is invaluable.

On the friends-side of this list Omar, Faisal and Aalia go first. Thank you for your advice, your patience, your faith and especially your understanding.

And finally, I want to mention the infinite gratitude I have and will always have for my family for being a fixed point in my life. Thank you for supporting me to pursue my curiosity in science.

Burooj Ghani
Hamburg

"Begin at the beginning," the King said, very gravely, "and go on till you come to the end: then stop."

Lewis Carroll,
Alice in Wonderland

Contents

Introduction 1

"Wir müssen wissen, wir werden wissen."

David Hilbert

1.1 Motivation

Acoustic signals are rich in information content. They provide engineering and scientific insights in myriad fields including underwater source localization, automated interpretation of human speech and animal vocalizations, acoustic monitoring, medicine, etc. For humans, the skill of listening to sounds and extracting relevant information comes effortlessly. However, for computers, this task is still quite challenging. Factors ranging from large variance in characteristics of a sound class, similarity in sound properties across classes, reverberation and noise, interference effects due to other signals, and overlapping sounds act as strong impediments when it comes to realizing the full potential of automated analysis of sound data. Nevertheless, as the amount of data grows, human effort required to manually process the acoustic data and recognize relevant patterns increases considerably. Furthermore, human cognition also places limits on identifying certain trends that may exist in

data but are hard for us to glean, especially at scale. Machine hearing entails developing computational methods that would enable us to capture the approximate statistical structure of acoustic signals. The more functional goal is to be able to analyze acoustic signals automatically to extract useful information that would assist in a wide range of applications.

One such utility is in the field of bioacoustics, and more specifically, bird vocalizations. Automated analysis of bird sounds can be utilized to answer some specific questions such as: Which bird species are vocalizing? When are birds present and vocalizing? How do different calls or songs relate to each other? What kind of calls or songs were produced? Are there differences in vocalizations between different sub-species of birds? (Bianco et al., 2019) Birds are relatively well studied since they are found in most environmental niches. This also makes avian population trends a useful indicator to measure the health of an ecosystem. The need for monitoring bird species populations has never been more dire owing to the stresses of human driven global warming that has accounted for a surge in global extinctions of animal populations in general and bird populations in particular. One in six bird species populations around the world are at a risk of extinction today as per the IUCN Red List (Ritchie and Roser, 2021).

Owing to the recent advances in recording and data storage technologies, affordable autonomous recording units can be deployed for extended time periods in favorable locations to collect huge amounts of data. This has led to a huge interest in developing state-of-the-art algorithms that can enable computers to analyze the data to automatically classify bird species based on their sounds. Compared to the advances on the hardware side, there is still an enormous dearth of computational methods that will allow for seamless analysis of huge amounts of acoustic data saving humongous time and effort that is spent in analysing the sound samples manually. Most of the literature available on automated recognition of bird species is either focused on specific bird species or the recordings are carefully chosen and of high quality (Priyadarshani et al., 2018). On the other hand, there are algorithms that are developed for general use based on very deep neural networks but these are extremely data hungry, have high computational complexity, and suffer from higher degrees of interpretability or adaptability issues (Purwins et al., 2019). Therefore, there is a need to build robust and scalable models that can be run on minimal computational budgets to carry out the task of automated analysis of unattended field recordings. While there have been significant advances in the recent past in the field of machine hearing, the field is still not as diverse and advanced as the field of machine vision and there is huge scope for improvement. The broad objective of this thesis is to build on the computational analysis tools that will assist automated monitoring of bird species based on their vocalizations.

1.2 Main Aims

1. Exploring performance of bird species classification by relying on simple, computationally inexpensive methods that can be trained on pre-computed sound features without a careful selection of bird species or recording samples.

2. Investigating the influence of number of selected bird species classes on classification performance.

3. Quantifying variations in songs within bird species using a purely algorithmic approach.

1.3 Thesis Outline

Chapter 2 starts off by describing the foundational concepts and tools from the areas of audio signal processing and machine learning that are necessary to grasp the workings of techniques and results presented in this thesis. This is followed by a recapitulation of requisite theoretical background on bird vocalizations (**Chapter 3**), motivation behind the impetus for analyzing bird sounds and presentation of a review of methods and techniques carried out in the past toward classification of bird species based on their sounds.

In **Chapter 4**, the utility of a simple statistical technique to perform the task of bird species classification is studied. Ensembles of cepstral coefficients are computed from bird vocalizations distributions are estimated from the time series of cepstral coefficients for different species. This is followed by comparing distributions of these ensembles by computing relative entropies to reveal group specificities. Next, a random forest classifier is employed and several experiments are conducted; to explore the predictive power of different descriptors, scanning for parameters to come up with the feature configuration that would provide the best performance, investigating the influence of segmenting signal parts from audio recordings, investigating the influence of bird sound type, and the influence of noise on the performance of classification. The goal here was to explore the complexities of the bird sound classification problem and understand the influence of different attributes on the overall performance of a classification system.

In **Chapter 5**, a shallow feed-forward neural network is fed with pre-computed sound features to study the robustness of bird sound classification. It is investigated in detail if and how classification results are dependent on the number of species and the selection of species in the subsets presented to the classifier. The number of species present in each subset is varied between 10 and 300, randomly drawing sounds of

species from a data set of 659 bird species to study the reliability of classification performance results.

In **Chapter 6**, the analysis is extended to explore the intra-species differences in bird species vocalizations in widespread species. Ornithologists have known that birds species vocalizations can vary even within the same species. Machine learning models are employed to classify vocal variation within widespread species in a way that does not require hundreds or thousands of hours of manual processing of recordings. Two widespread bird species, House Wren and Yellowhammer, are considered in this work to explore the possible variations in songs in case of the former and classification of various established dialects in case of the latter species. In contrast to existing approaches these issues are approached in a purely algorithmic manner using the classification approach developed in the previous chapters. Our motivations to choose a shallow feed-forward neural network and a random forest classifier over very deep networks lie mainly in the model simplicity, the lower computational costs, and the relatively small amount of data required to train such networks. We wanted to analyze the classification performance using a simple model that can be trained with pre-computed sound features.

In **Chapter 7**, the findings of different computational experiments carried out in this thesis are summarized and a brief outlook for future work is presented.

Chapter	Title	Journal	Status
5	A randomized bag-of-birds approach to study robustness of automated audio based bird species classification.	Applied Sciences	Published
6	Quantifying variability in songs for House Wren species.	Ecology and Evolution	In preparation
6	Automated Classification of dialects of Yellowhammer species.	Ecological Informatics	In preparation

Table 1.1: Manuscripts from analysis in this thesis.

Machine Hearing Foundations 2

"If you want to find the secrets
of the universe,
think in terms of energy,
frequency and vibration."

Nikola Tesla

2.1 From Sounds to Acoustic Signals

Sound can be defined as the auditory perception of pressure waves that are generated by an oscillating object. Waves travel in the form of compressions and rarefactions of particles of an elastic medium such as water or air. These compressions and rarefactions are alternating regions of high and low pressure respectively. The particles retain their equilibrium position while transporting energy in the form of pressure waves that propagate through the medium. The number of times the object producing the waves oscillates back and forth in one second gives the sound its key characteristic known as frequency. Frequency of oscillation is measured in cycles per second or Hertz (Hz). It is the physical basis for our sensation of pitch. Humans can detect sounds in the stated frequency range between 20 Hz and 20 kHz.

Acoustic transducers, also known as microphones[1], are artificial sensors used to

[1] or hydrophones if you are working under water.

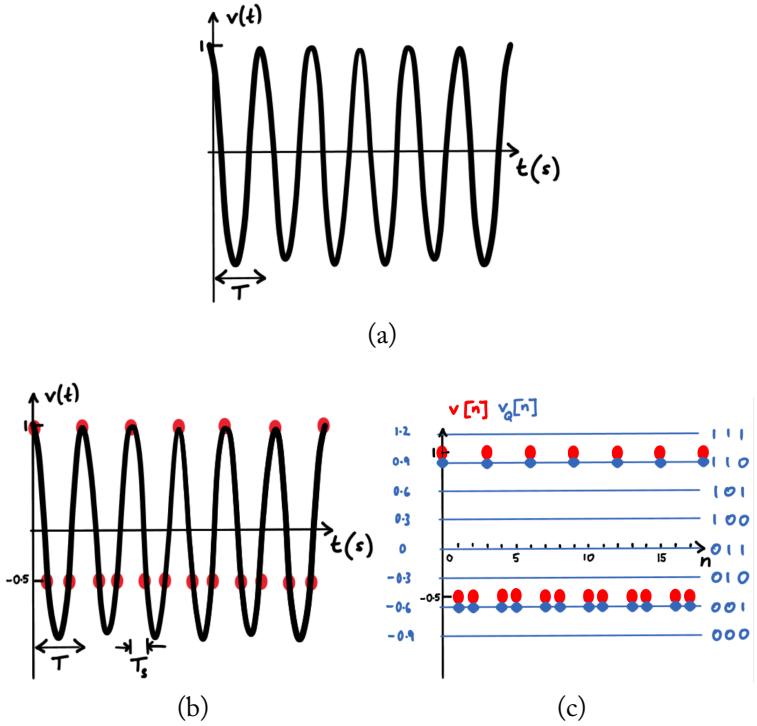


Figure 2.1: Schematic for analog to digital conversion. a) Analog Signal: An analog signal $v(t)$ where T denotes the period i.e., the duration of a full oscillation. b) Sampling: The analog signal $v(t)$ is uniformly sampled 3 times within each period T to produce a time-discrete sequence of samples v[n]. c) Quantization: A 3-bit quantizer with 8 quantization levels that are uniformly spaced is applied to the samples in v[n] to generate a fully digital signal. This results in each sample in $v[n]$ being approximated to the nearest level of the quantizer. Figure adapted from Varodayan (2014)

mimic our auditory perception. These devices are placed in the sound field to record sounds and convert sound pressure difference over time into a time varying voltage difference to provide an electrical representation of sound. These continuous time and pressure variations of electrical signals are known as acoustic signals *v(t)*. In order to store these representations on a computer, the continuous acoustic signals are converted to a digital format *v[n]*. This process is called analog to digital conversion. The digitization process primarily includes two processes to reduce the resolution of continuous signals while maintaining the information content that allows us to perceive the sound. The two processes are sampling and quantization. Sampling, also referred to as discretization, entails sampling the continuous acoustic signal at discrete intervals in time, usually spaced linearly. Sampling frequency (f_s) determines the number of samples that are measured from the continuous acoustic signal. Consequently, a digital recording is composed of a series of sample values proportional to

the original sound pressure at discrete time intervals given by the sampling frequency. The sampling frequency is estimated using *Nyquist–Shannon sampling theorem* which establishes a relation between the sampling frequency and the range of recorded frequencies in the continuous acoustic signal. To elaborate, it states that a digital audio signal can only capture frequencies less than half the sampling frequency. Following this criterion, the analog signals are sampled at a sampling frequency of at least two times the highest recorded frequency or the maximum frequency relevant for us ($f_s = 2f_m$ where f_m is the highest relevant frequency).

The other process that assists analog to digital conversion is called quantization. Quantization is sampling performed in the pressure or amplitude domain as opposed to discretization which is performed in the time domain. The idea is to approximate a continuous pressure value from a set of predefined amplitude values. These predefined values depend on a parameter called bit depth which is the number of bits or binary digits used to specify the amplitude. Consequently, the bit depth places a limit on the number of discrete values that can be used to represent the amplitude at a specific time interval. In more detail, 2^n different amplitude values can be represented by a bit depth of n bits. Fig. 2.1 graphically illustrates the digitization process.

2.2 Time-Frequency Representation

It was discussed in the last section how frequency is one of the key characteristics of sound signals. It determines the number of times the pressure oscillates in one second. While this is true for a signal that is composed of a single frequency, real world sounds are more complex and contain more than one frequency component at any moment in time. Also, it is quite useful for many real-world applications to have the frequency distribution information of sound signals. Fig. 2.2 illustrates the idea further. Therefore, the central question at this point is how can we take such a signal and decompose it into its constituent frequencies. The *discrete Fourier Transform* (DFT) is one such technique that can be used to transform an acoustic signal in time domain into its component frequencies along with corresponding amplitudes and phases for each frequency component. Such a distribution of amplitudes against the frequency components of an acoustic signal is known as its frequency spectrum. More formally, the discrete Fourier Transform is defined by:

$$\mathbf{F}[k] = \sum_{n=0}^{N-1} \mathbf{v}[n] e^{-2\pi i k n/N}, \quad k = 0, 1, \ldots, N-1, \tag{2.1}$$

where, $\mathbf{F}[k]$ is the DFT of the acoustic signal sequence $\mathbf{v}[n]$. $\mathbf{F} = \mathbf{F}[0], \mathbf{F}[1], \ldots, \mathbf{F}[N-1]$ are the fourier components, see e.g., (Osgood, 2007).

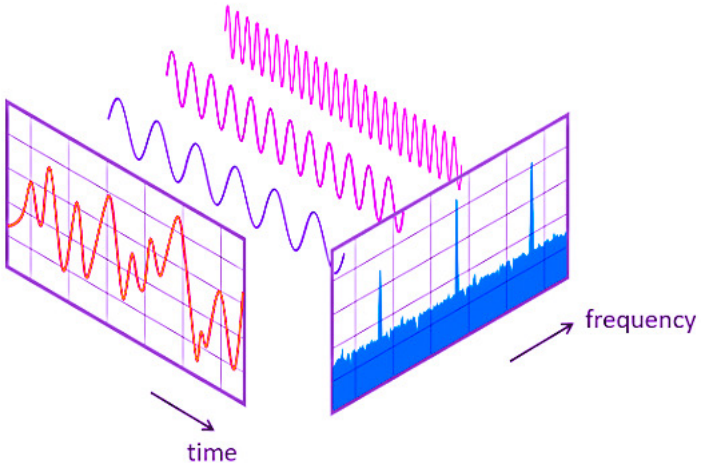


Figure 2.2: A signal in time and frequency domain. Image reproduced from nti audio

One way of understanding what is happening in Eq. 2.1 is in terms of correlation between sinusoids of different frequencies and the acoustic signal. In order to understand this further, let us use *Euler's formula* [2] to break the exponential in the above equation into its sine and cosine components where $\theta = 2\pi kn/N$. We can hence re-write Eq. 2.1 as:

$$\mathbf{F}[k] = \sum_{n=0}^{N-1} \mathbf{v}[n](\cos(2\pi kn/N) - i\sin(2\pi kn/N)), \tag{2.2}$$

and consequently:

$$\mathbf{F}[k] = \sum_{n=0}^{N-1} \mathbf{v}[n]\cos\left(\frac{2\pi kn}{N}\right) - i\sum_{n=0}^{N-1} \mathbf{v}[n]\sin\left(\frac{2\pi kn}{N}\right) \tag{2.3}$$

We can see from Eq. 2.3 that both summations are basically correlations between the signal ($\mathbf{v}[n]$) and a sinusoidal function. This calculation is repeated for different values of k which gives us fourier coefficients for different frequencies. From the cosine part we derive the real part that gives us the amplitude for a particular frequency range present in the signal while the sine part which is imaginary gives us the corresponding phase. In this way, we generate a spectrum of a signal from its amplitude values. While the frequency spectrum provides valuable insight about the distribution of frequencies in a signal, it is not helpful if we are interested in time localizing various frequencies in a signal. Frequency spectrum provides an averaged estimate of

[2] $e^{-i\theta} = \cos\theta - i\sin\theta$

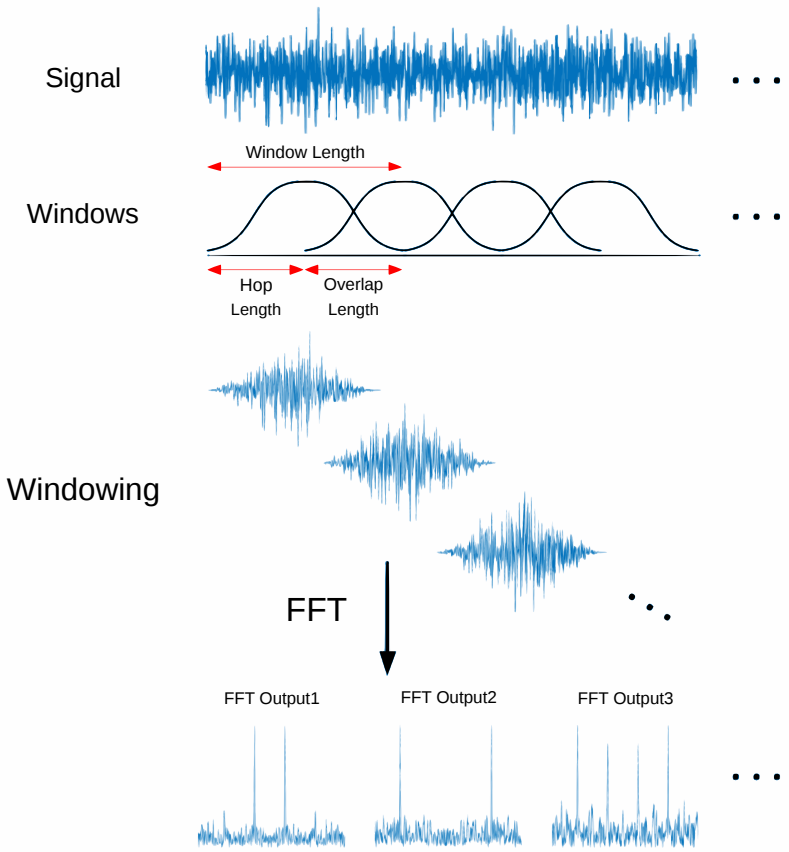


FIGURE 2.3: Overview of short term Fourier Transform. Image adapted from Jeon et al. (2020)

different frequencies in the entire signal. In order to recover the time information, Gabor introduced *short-time Fourier transform* (STFT). The simple idea is to divide the long signal into small consecutive segments and then perform fourier transform on these individual frames separately in order to understand how various frequencies evolve. The kth coefficient of DFT for the tth time frame of $\mathbf{v}[n]$ is computed as:

$$\mathbf{F}[t,k] = \sum_{n=0}^{N-1} w[n]\mathbf{v}[tH+n]e^{-\frac{-i2\pi kn}{N}}, \tag{2.4}$$

where $w[n]$ is the window function that is used to make sure frames on which the DFT is computed are periodic and to account for the discontinues at the end points that arise due to non-integer number of periods filling the frames. There are several types of windowing functions, the choice of which is contingent on different applications. Some commonly used windowing functions are Rectangular, Hamming, Hanning, and Blackman-Harris. H is another parameter which is referred to as hop

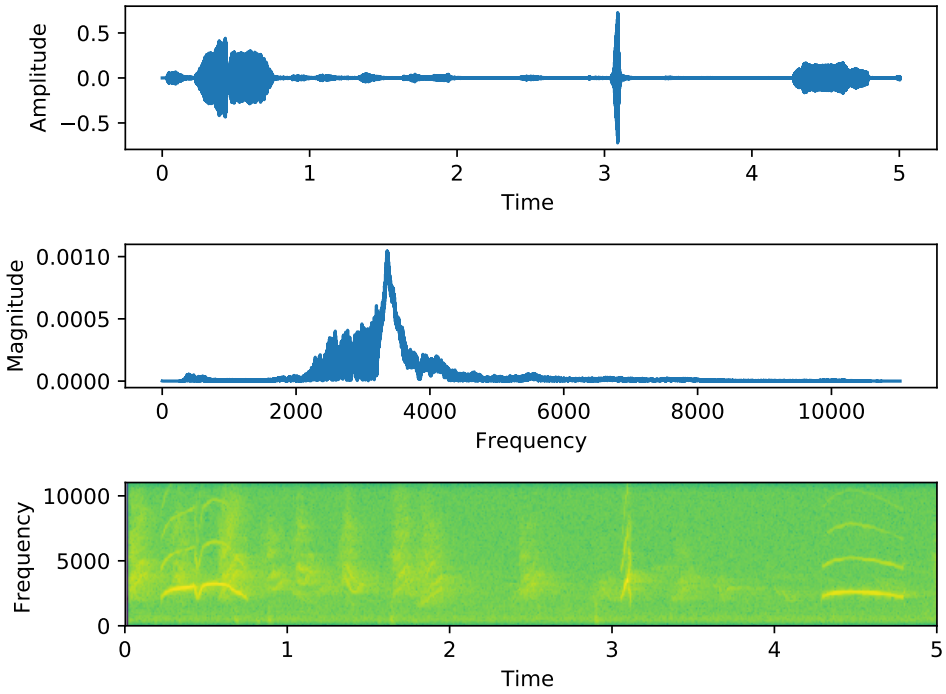


Figure 2.4: Time domain, frequency domain, and time-frequency domain representation of a sound signal of a bird song of 5s duration. a) Waveform of a 5s signal b) A time averaged frequency spectrum of the signal c) Spectrogram representation that shows frequency distribution over time.

size and is specified in number of samples. Hop size determines the fraction of overlap between consecutive frames. It is common practice to introduce an overlap by choosing hop size smaller than the length of the frame, this leads to smooth STFT representations and also allows statistical dependencies between time frames (Virtanen et al., 2018).

Spectrograms provide a way of visualizing various frequency contributions of a signal as it evolves over time, see e.g., (Müller, 2015; Smith, 2007). Fig. 2.4 shows a representation of a big song signal along with its spectrum and spectogram. It is evident from the figure that, while the frequency spectrum provides the frequency distribution of the entire signal, the spectrogram informs on an elaborate frequency evolution along the time horizon.

2.3 Sound Scene Analysis: Machine Hearing

Acoustic signals are rich in information content. For humans, the skill of listening to sounds and extracting relevant information from them comes effortlessly. This ability

of the human auditory system to extract separate sound events from complex sound mixtures is phenomenal. However, for computers, this task is quite challenging. *Machine Hearing* entails developing computational methods to capture the approximate statistical structure of acoustic signals (Mobin, 2019). The more functional goal is to be able to analyze acoustic signals automatically to extract useful information. In general, most machine hearing tasks can be broadly divided into two main categories: Classification and Detection. Classification deals with assigning an acoustic signal to one of the several predefined classes. For instance, different classes could be several bird species and the goal of the classifier would be to analyze the acoustic signal to assign one bird species label to it. When the number of classes is limited to two, it is termed as binary classification, and when the number of categories is more than two, it is called multi-class classification. Detection algorithms, on the other hand, aim at time localizing different events or sources of sounds within an audio recording.

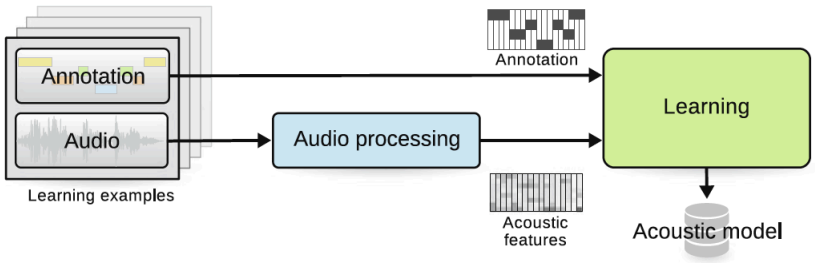


FIGURE 2.5: Overview of machine hearing process. Figure reproduced from Virtanen et al. (2018).

A typical machine hearing system involves acquiring acoustic signal, pre-processing the signal, extracting features, and decision-making (Li and Cox, 2019). An acoustic signal is recorded using a transducer such as microphone and stored in a digital form. The stored signal is then pre-processed. This may involve several processes such as denoising, frequency filtering, pre-emphasis, sound-source separation, peak normalization, re-sampling, etc. The goal of this step is to enhance specific properties of the acquired signal that best assists the overall performance of the task the system aims at solving. The next step entails extracting audio features from time series of audio signals. Extracting features allows obtaining lower-dimensional compact statistical representations while preserving the distinguishing characteristics of the signal in a non-redundant manner. In addition to reducing the computational costs, feature extraction maximizes classification performance of the system. Studies have shown that models with aggregated features achieve a better classification performance compared to a single feature (Xie and Zhu, 2019). The final decision-making step can range between something as simple as setting a threshold on a uni-

dimensional feature to quantifying similarity or dissimilarity involving a set of features in a higher dimensional space. The decision making algorithms grow in complexity as acoustic signals become more complex. Machine learning algorithms are commonly used to handle the demands of sophistication (Li and Cox, 2019). The simple idea of machine learning is to use large amounts of data to learn models that can capture necessary statistical properties inherent in the data. Lately, *Artificial Neural Networks* (ANNs) have proven to be one of the most effective machine learning frameworks when it comes to learning representations of sounds and images and have produced state-of-the-art results (Mobin, 2019; LeCun et al., 2015). These work on the principle of supervised machine learning where a set of input-output pairs are introduced to the system and the aim is to learn the mapping of inputs to the outputs. In case of acoustic signals, we have acoustic features extracted from a signal to be analyzed and reference annotations provided for each signal respectively. Fig. 2.5 provides a schematic of the overall machine learning process for acoustic analysis.

2.3.1 Machine Learning

In this section, the different machine learning techniques used in this thesis are introduced. These techniques are commonly used and widely known by the machine hearing community. A brief overview is presented here.

Decision Trees

Decision trees can be applied in both classification and regression aspects (Breiman et al., 2017). The core idea behind decision trees is to split or stratify the input space into sub-spaces. The sub-space regions are defined by a set of binary splitting rules based on which quantitative or qualitative decisions are made. Stratification of the input space can be nicely summarized in a decision tree and that makes these models very insightful data analysis tools. The biggest advantage of decision trees is that they are simple algorithms, easily interpretable, and nice to visualize. In order to make binary splits, the classification error rate is used as a criterion. It gives the fraction of training samples in a region that do not belong to the most commonly occurring class of training samples in that region and is defined by:

$$E = 1 - \max_{k} \left(\hat{p}_{rk} \right), \tag{2.5}$$

where $\hat{p}_{rk}$ gives the proportion of training samples in the r^{th} region that belong to the k^{th} class. In practice, another criterion known as the Gini index is mostly used since the classification error has been found to not be sufficiently sensitive to grow trees, see e.g., (James et al., 2013). The Gini index provides a measure of total variance

across K classes and is given by:

$$G_r = \sum_{k=1}^{K} \hat{p}_{rk} \left(1 - \hat{p}_{rk}\right). \tag{2.6}$$

Note that G_r is 0 when a sub-space has samples from only one class. The Gini index has the property of encouraging purity in a region i.e., favours regions with high proportions of data samples that purely belong to one of the classes, since the goal is to maximize homogeneity in a region. The tree is then recursively split to minimize:

$$\min \sum_{r=1}^{R} G_r \tag{2.7}$$

Random Forests

Although decision trees are very simple and therefore easy to interpret, they usually suffer from sub optimal performance. The variance across different classes is high. Therefore, if the training samples are randomly split into two parts and decision trees are fit to both halves, the results for the two could be significantly different. In order for a model to be robust in terms of predictive power, it should deliver similar results on different datasets i.e., the variance should be low. Bagging or bootstrap aggregation provides a solution to overcome the problem of low variance. In this procedure, B number of decisions trees are trained using B separate training datasets. In case, the data is limited, the same can be done using bootstrapping, where repeated samples are drawn B times, with replacement, from the same data set to train B number of decision trees. Consequently, for a given test sample, the majority vote from the predictions made by the B trees for that sample determines the predicted class for the sample.

The problem with such bagged trees is that they can be highly correlated. If there are a few features in the data set that are very discriminative, all bagged trees will use these features in the top splits to make the first decisions and therefore are going to be very much the same in all trees. Random forests provides a solution to this problem (Breiman, 2001). In addition to building an ensemble of decision trees by bootstraping the data set, a procedure called feature bagging is employed. For each decision tree, every time a split is considered out of a total of p features, a random selection of m features are chosen as split candidates. At each split, a fresh sample of m features is taken, where m is typically set to be equal to $\sqrt{p}$. This makes sure that each tree that is built is different based on the random selection of features and the result is an ensemble of trees that are uncorrelated (James et al., 2013). Random forests have shown very strong performance in a very wide range of empirical

evaluations compared to several other supervised learning algorithms (Caruana and Niculescu-Mizil, 2006).

Artificial Neural Networks

Neural networks derive their name from the fact that they were initially developed to model the human brain. These models are sometimes also referred to as single hidden layer back-propagation networks. A neural network is a two-stage non-linear statistical model that can perform both classification and regression (Friedman et al., 2001). Fig. 2.6 provides a graphical representation of a typical neural network. These models constitute a sequence of layers where each layer is an affine transformation followed by a non-linear transfer function σ:

$$f_i(\mathbf{x} \mid \boldsymbol{\theta}) := \sigma_i \left(\mathbf{w}_i^\top \mathbf{x} + \mathbf{b}_i \right),$$

$\boldsymbol{\theta} = (\mathbf{w}_i, \mathbf{b}_i, \sigma_i)$ constitutes the parameter space where $\mathbf{w}_i$ are weights, $\mathbf{b}_i$ are biases and σ_i are transfer functions for different layers i. The transfer function σ_i is usually selected as the sigmoid function $1/(1 + e^{-v})$. The values Z_m of the units in the

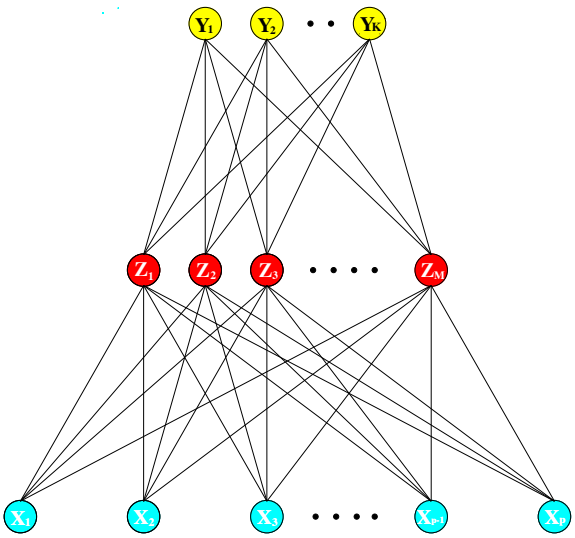


Figure 2.6: Schematic of a feed forward neural network with a single hidden layer (Friedman et al., 2001).

middle are not directly observed and therefore are referred to as the hidden units. These values are sometimes called derived features and are computed from linear combinations of input data. There can be, in general, more than one hidden layer in a neural network. The goal of the neural network is to learn the value of parameters $\boldsymbol{\theta}$ that generates the best function approximation for the training data (Virtanen et al.,

2018; Goodfellow et al., 2016). For classification, cross-entropy is usually used as an error function to measure the model fit. Gradient descent is used as a generic method to minimize cross-entropy loss.

2.3.2 Commonly Used Acoustic Features[3]

The acoustic features that have been employed in different experiments in this work are described as follows:

Mel Spectra

In Section 2.2, it was discussed how, in order to capture the time evolution of a signal, an STFT is computed on short time consecutive frames. Spectrum computed using STFT provides the power spectrum for these successive frames of the audio signal. The frequency representation thus produced is linear. The inception of the Mel scale is inspired by human auditory perception (Stevens et al., 1937). To elaborate, the idea is that the human auditory system perceives both magnitudes and frequencies of individual frequency components in a highly non-linear manner. Following this understanding, a new type of spectral features were introduced, that are less discriminative of higher frequencies and more discriminative of lower frequencies. This is achieved by allowing larger bandwidths at higher frequencies and narrow bandwidths at the low frequencies (Virtanen et al., 2018). Mel-scaling is implemented using filter banks in which triangular filters are applied to the power spectrum on non-linearly spaced frequency ranges. Fig. 2.7 shows a Mel-spaced filterbank with 20 triangular filters. To compute filterbank energies, each filterbank is multiplied with the corresponding power spectrum and then the coefficients are added up to generate a Mel spectra.

Cepstral Coefficients

It is well understood that the transformation from time domain to frequency domain allows for conversion of the convolution of two discrete time series to a multiplication of the two sequences. If a logarithm of the spectrum is followed by an inverse transform, it turns out that this also provides useful information utilizing the property of logarithms to convert products into sums. The relevance of this concept derives from the idea that an audio signal recorded in a noisy reverberant environment may be written as a convolution of the original audio signal and the impulse response of the environment (Vary and Martin, 2006). Therefore, if the recorded signal has a well-defined spectral vs excitation envelope, a cepstrum allows for a separation of those

[3] **Some parts of this section have been published in (Ghani and Hallerberg, 2021)**

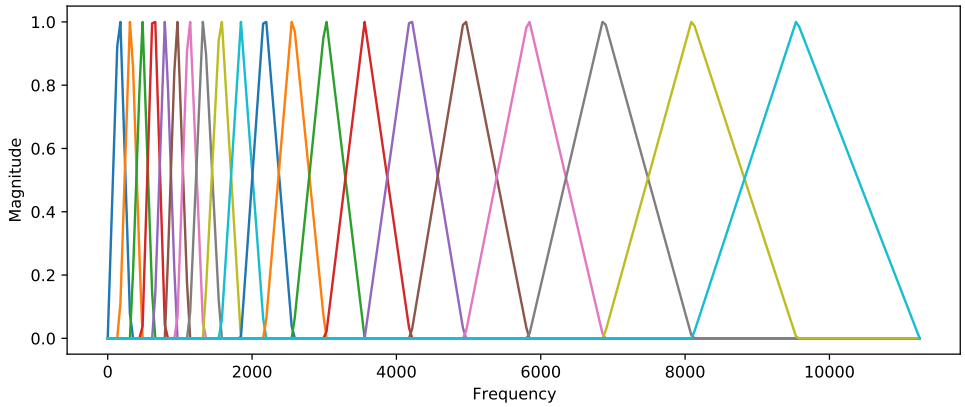


Figure 2.7: Mel-filterbank containing 20 filters that projects STFT bins onto Mel-frequency bins. The minimum frequency is 0 Hz and the highest frequency is 11250 Hz.

parts. The cepstrum of a signal $v[n]$ can be computed as:

$$C(v[n]) = F^{-1}[\log(|F[v[n]]|^2)], \tag{2.8}$$

which entails computing the Fourier transform of the signal, then taking a magnitude squared. By taking the magnitude we throw away the phase information. This is followed by taking a logarithm and then an inverse transform to get a cepstrum $C(v[n])$ of the signal $v[n]$. By taking an inverse transform what we get is something like an inverted frequency which is termed as quefrency. This can be thought of as the frequencies of the frequencies generated by the first Fourier transform. The lower coefficients in the cepstrum will represent the slowly changing spectral information that corresponds to the spectral envelope and is the useful part of the signal like information about formants, the relative phonemes, the timbre etc. As we move towards higher quefrency values in the cepstrum, we will get spectral details or fast changing values in the spectrum. Consequently, when we move from spectral to cepstral domain, we can achieve a separation between the spectral envelope and spectral details that correspond to the representation of the impulse response of the environment.

Mel-Frequency Cepstral Coefficients (MFCCs)

The coefficients of the Mel spectra obtained by applying a Mel-spaced filterbank to the power spectrum of a signal turn out to be highly correlated which is not desirable for some machine learning algorithms such as Random Forests. In order to decorrelate the coefficients of the mel spectra, they are first log transformed and then a *discrete Cosine Transform* (DCT) is applied to obtain *Mel-Frequency Cepstral Co-*

efficients (MFCCs). This also produces a compressed representation (Logan et al., 2000). For a typical classification task, only the first 10-20 coefficients are retained while the rest are discarded. This follows the same reasoning as the linear cepstral coefficients, that the latter coefficients represent fast changes in the mel spectra that do not contribute towards automatic recognition of desired patterns.

Deltas

The cepstral coefficients or the MFCCs thus computed describe the power spectral envelope of different frames of the acoustic signal, but some audio signals have also been shown to carry useful information in the trajectories of coefficients over time. The delta coefficients are therefore computed to encapsulate such information and can be defined as:

$$d_n = f_n - f_{n-1} \tag{2.9}$$

where d_n gives the delta coefficient for a feature f at time instant n. It is a common practice, for instance, in the speech recognition tasks, to append 10-20 delta coefficients to the MFCCs which then forms an input to the machine learning algorithm.

Spectral Centroid

The spectral centroid measures the frequency where energy of a spectrum is centered. In other words, it localizes the center of mass of the spectrum and is calculated as a weighted mean of the frequencies the signal is composed of:

$$s_c = \frac{\sum_k S(k) f(k)}{\sum_k S(k)}, \tag{2.10}$$

where, S(k) is the spectral magnitude at frequency bin k and f(k) represents the center frequency of the bin (Klapuri and Davy, 2007).

Spectral Rolloff

The spectral rolloff gives the frequency f(k) below which a predefined percentage (usually set to 0.85) of the total spectral energy is concentrated (Smith, accessed <2019>).

Zero-Crossing Rate

The zero-crossing rate r measures the smoothness of a signal. It is the rate at which the signal changes its sign from negative to positive or vice versa (Giannakopoulos and Pikrakis, 2014).

The Root-Mean-Square Energy

The root-mean-square energy of a signal gives the signal's total energy and is defined as:

$$e_{rms} = \sqrt{2 \sum_{k} |S(k)|^2} \tag{2.11}$$

where $S(k)$ is the spectral magnitude at frequency bin k (McFee et al., 2019).

Spectral Bandwith

Spectral Bandwith measures if the power spectrum is concentrated around the spectral centroid or spread across the spectrum. It is computed as:

$$s_b = \sqrt{\frac{\sum_k (k - s_c)^2 \cdot |S(k)|^2}{\sum_k |S(k)|^2}} \tag{2.12}$$

where s_c is the spectral centroid, $S(k)$ is the spectral magnitude at frequency bin k (Abreha, 2014).

2.3.3 Measuring Classification Success[4]

This section introduces indices that have been used to measure classification performance in different experiments (Sunasra, 2019; Grandini et al., 2020). Each experiment involves n species and a data set of a total of $4m$ sound samples. To elaborate, $3m$ samples are used for training, whereas the remaining m samples are employed to evaluate the quality of classifications. For each species $j = 1, 2, ..., n$ a data set of $4m$ samples is processed. These evaluations are then done by computing a confusion matrix for each species j, containing:

$c_{tp}(j)$ the number of classifications which are true positives,
$c_{tn}(j)$ the number of classifications which are true negatives,
$c_{fp}(j)$ the number of classifications which are false positives,
$c_{fn}(j)$ the number of classifications which are false negatives.

The resulting elements of the confusion matrix enter into a computation of more advanced metrics for measuring classification success such as precision, recall, accuracy, mean-average-precision and receiver-operator characteristics. To understand how the entries of each species' confusion matrix influence the outcomes of these summarizing measures, more detailed descriptions of their computations are introduced as follows:

[4] **Some parts of this section have been published in (Ghani and Hallerberg, 2021)**

Confusion Matrix

A confusion matrix is an $N \times N$ matrix, where N is the number of classes that is used to evaluate performance of a classification model. The columns of the matrix represent the true values while the rows represent the predicted values of the target variable that can either be true or false.

Precision

The metric gives the measure of reliability of our predictions. The formula to compute precision for a bird species j is

$$p(j) := \frac{c_{tp}(j)}{c_{tp}(j) + c_{fp}(j)}. \tag{2.13}$$

Therefore, precision for a species j indicates how many true positives the model predicted out of all positives. Therefore, higher the precision, the more confident a model is about its predictions. In order to compute precision for entire test data set, we average over all species

$$P = \frac{1}{n} \sum_{j=1}^{n} p(j). \tag{2.14}$$

Recall

This metric gives the measure of predictive power of a model. The formula to compute recall for each bird species j is

$$r(j) := \frac{c_{tp}(j)}{c_{tp}(j) + c_{fn}(j)}. \tag{2.15}$$

Therefore, recall for a class species j would mean, of all actual positives in the test data set, how many did the model predict as positive. Therefore, the higher the recall, the more positive samples model correctly classified as positive. In order to compute recall for the entire test data set, we average over all species:

$$R = \frac{1}{n} \sum_{j=1}^{n} r(j). \tag{2.16}$$

Accuracy

While precision and recall are computed for each class separately in a multi-class classification problem, the accuracy A is computed for the entire test data set using

$$A := \frac{\sum_{j=1}^{n} c_{tp}(j)}{m \cdot n}. \tag{2.17}$$

So, out of all test samples how many were correctly classified.

Area Under ROC Curve (AUC)

An *Receiver Operating Characteristics* (ROC) curve shows performance of a classification model at different classification thresholds. The curve is computed by plotting true positive rate (r_{tp}) against false positive rate (r_{fp}) at these thresholds. The true positive rate for a bird species j is defined as:

$$r_{tp}(j,\rho) := \frac{c_{tp}(j,\rho)}{c_{tp}(j,\rho) + c_{fn}(j,\rho)}, \tag{2.18}$$

and the false positive rate for a bird species j is defined as

$$r_{fp}(j,\rho) := \frac{c_{fp}(j,\rho)}{c_{fp}(j,\rho) + c_{tn}(j,\rho)}, \tag{2.19}$$

with ρ denoting a probability threshold that is varied from 0 to 1 in order to obtain the ROC curve. *Area Under ROC Curve* (AUC) gives an aggregate measure of classification performance. The ROC was originally developed for a binary classifier and has later been generalized for multi-class classification system (Hand and Till, 2001). Test set labels are binarized by employing either the one-vs-one or the one-vs-rest configuration. We have employed the one-vs-one configuration for our task. To elaborate, different sound samples are ranked by their probabilities and then false positive and true positive rates are computed by choosing different probability cut-offs ρ to generate the ROC curve. AUC is computed as the area under the ROC curve. In the end, an average across species is computed to get one AUC value for the entire data set, i.e.,

$$AUC = \frac{1}{n} \sum_{\rho} \sum_{j=1}^{n} r_{tp}(j,\rho). \tag{2.20}$$

Mean Average Precision (mAP)

Mean Average Precision (mAP) gives us a way of characterizing the performance of a classifier by monitoring how precision changes on varying the classification probability threshold, one the model uses to make a decision if a bird sound sample belongs to a class j. A good classifier will maintain a high precision as recall increases while a poor classifier will take a hit on precision as recall increases with changes in threshold. To elaborate further, to compute Average Precision for a species j, a list of probabilities is generated in which the discrimination probabilities our model has assigned to all the test samples for class j are stored. The list is then sorted by decreasing

probabilities and each element is assigned a rank l. By varying the rank l (by gradually lowering the probability threshold), a list of true positives and false positives is generated. Note that, as the classification threshold is lowered, the model labels increasingly more samples as positive. This will lead to an increase in false positives. The list is consequently employed to come up with a list of precision values at different ranks $p(l)$. Considering all the L cases in the list where the sound sample belongs to class j, the average precision is computed as:

$$P_A(j) := \frac{\sum_{l=1}^{L} p(l)\mathbf{1}(l)}{c_{tp}(j)}, \tag{2.21}$$

where $\mathbf{1}(l)$ is an indicator function that equals unity if the sample at threshold l is a true positive. The mean average precision P_{mA} is then computed by averaging over all classes (species) (Kahl et al., 2019).

$$P_{mA} := \frac{\sum_{j=1}^{n} P_A(j)}{n}. \tag{2.22}$$

2.3.4 Basic Statistics

Audio signals are time varying phenomena and come with intrinsic randomness due to the background noise and/or interference effects added to them en-route. They can therefore be modeled as discrete stochastic processes. A discrete time random process is defined as a sequence of random variables $X(n)$, where,

$$n = 0, 1, 2, \ldots, \tag{2.23}$$

is the time index as the signal evolves, see e.g., (Gray and Davisson, 2004). In order to extract the statistical structure of a random signal at a certain time step, one would need to have access to an ensemble of $X(n)$ sequences. However, in reality, it is not always feasible to observe more than one ensemble sequence of a random process. In case of stationary signals, ensemble averaging can be replaced by the time averaging. In case of non-stationary signals, for instance bird vocalizations, the signal is sliced into short time analysis frames. The features are extracted from these analysis frames that are assumed to be in a quasi-stationary state (Virtanen et al., 2018). Based on the values of features for different short time analysis frames, summary statistics are computed for each feature component to provide an estimate of the structure of the assumed underlying distributions representing different bird vocalizations. The sample moments that have been employed in this work are briefly defined as follows:

Mean

Sample mean of a distribution is the estimation of the first moment of the random variable X and is defined as,

$$\hat{\mu} = \overline{\mathrm{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{X_i}, \tag{2.24}$$

and provides an estimate of the average value of data.

Variance

Sample variance of a distribution is the estimation of the second moment of X and is defined as,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \hat{\mu}\right)^2, \tag{2.25}$$

and provides a measure of the spread of the distribution about the mean.

Skewness

Sample skewness of a distribution is the estimation of the third moment of the random variable X and is defined as,

$$\hat{\mu}^3 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^3, \tag{2.26}$$

and provides a measure of asymmetry of distribution about its mean. It can take all values on a real number line. A negative skewness would mean that the values are spread out more to the left of the mean and similarly, a positive skewness would mean the values are spread more to the right.

Kurtosis

Sample kurtosis of a distribution is the estimation of the fourth moment of X and is defined as,

$$\hat{\mu}^4 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^4, \tag{2.27}$$

and provides a measure of tailedness of a distribution and therefore quantifies to what degree outliers are present in the data.

Bird Vocalizations 3

"A bird doesn't sing because it
has an answer,
it sings because it has a song. "

Maya Angelou

3.1 Bird Sounds

Birds, in a way similar to humans and many other vertebrates, produce sounds by making use of their respiratory system. The quasi-periodic flow of air enables the respiratory muscles to transfer the energy to a small structure that then generates sound. While it is *larynx* in case of humans that is responsible for sound production, in birds, the vocal organ that produces sound is called *syrinx*. It is located at the base of trachea in contrast to the larynx that lies at the top of the trachea. To elaborate further, air propagates from the lungs through the bronchi into the syrinx which is where the vibrations get produced. The sound before emerging out as vocalizations gets modulated while passing through the trachea, the larynx, the mouth, and finally the beak of the bird, see e.g., (Bastas, 2012; Fagerlund, 2004; Catchpole and Slater, 2003). Although this is a simplified understanding of the sound production mechanism in birds, it captures the overall idea to assist the flow of details that will follow. The capability of producing sounds provides birds with an efficient means to com-

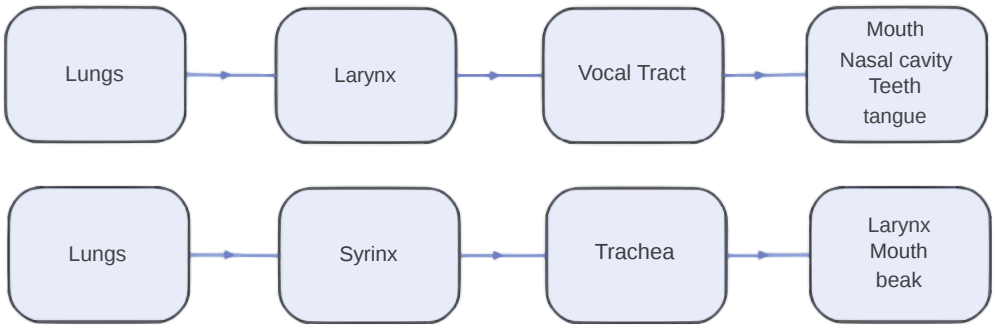


Figure 3.1: Human and bird sound production models (top and bottom, respectively). The figure has been reproduced from Potamitis et al. (2014)

municate as sound can traverse long distances, does not require visual contact, and is not hindered by low visibility for instance during nighttime, bad weather conditions or dense environments, see e.g., (Potamitis et al., 2014). Bird vocalizations that are produced by the vocal organ of birds can mainly be divided into *calls* and *songs*. Calls are usually classified as sounds that are shorter and less complex. These sounds are used in different contexts such as to share information about food sources, to facilitate contact between parents and their young offspring, to maintain flock cohesion, to signal alarm, etc. Songs, on the other hand, take the form of more complex vocalizations that are used mainly in the context of mating (Ball and Hulse, 1998). While calls are produced by all birds, only a subset of birds also produce songs.

It has historically been held that bird song is exclusively a male trait (Ball and Hulse, 1998). This observation was also fundamental to the formulation of Darwin's theory of sexual selection (Darwin, 1872). The female song instances were dismissed as atypical, scarce, or outcome of hormonal aberrations (Catchpole and Slater, 2003). However, more recent studies have reversed this classical assumption and there has been a growing acknowledgment in the past decades that female bird songs are also common, especially in the tropical regions (Odom et al., 2014). Fig. 6.1 shows spectrogram representations of calls and songs for three different bird species that have been recorded in Germany.

Songs are often simple harmonics that can contain single sinusoidal traces, overtones, and stronger formants. Due to their distinctive and elaborate nature, songs also allow for a greater possibility for automated sound-based bird species classification (Stowell and Plumbley, 2010). Birds can produce more than one version of their species' songs and these different versions are called song types. Bird songs can be hierarchically broken down into different sections. Each song consists of a series of phrases that occur in a certain pattern. The phrases in turn are composed

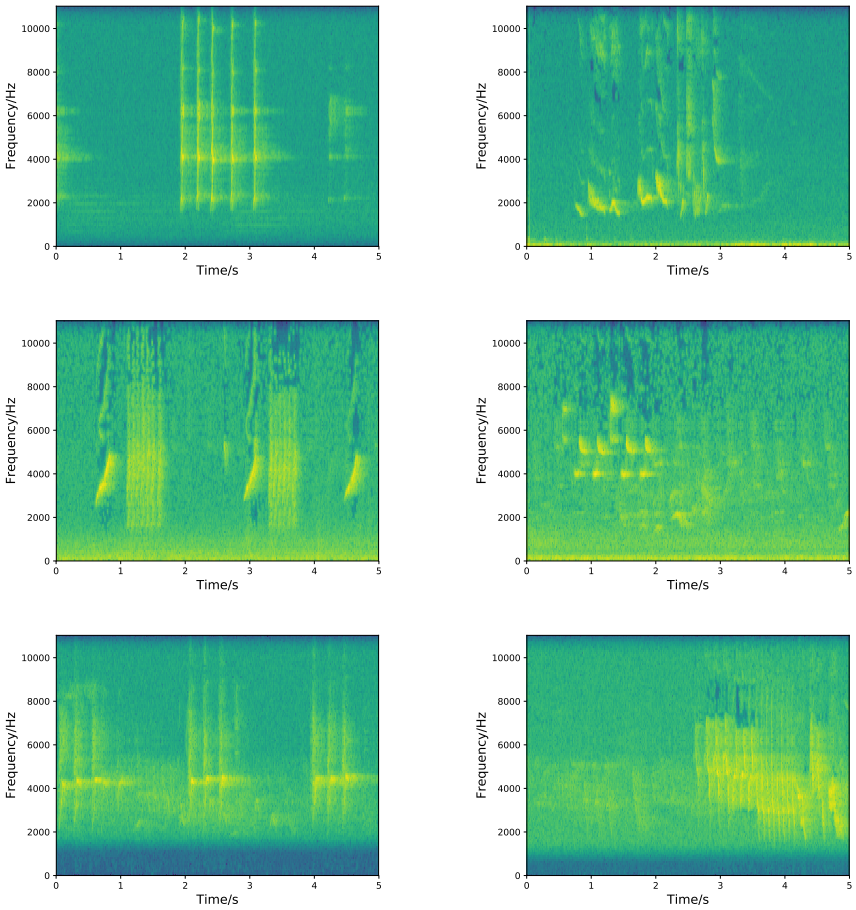


FIGURE 3.2: Spectrograms of bird sounds for three species recorded in different locations in Germany. The left column corresponds to calls and the right column corresponds to songs. In the top row is the bird species Common Blackbird (*Turdus merula*). The call is recorded in Essen-Werden, Nordrhein-Westfalen while the song is recorded at the University of Göttingen[1]. The center row is the species Great tit (*Parus Major*). Both the call and song for this species are recorded in the Hannover region (near Uetze). The bottom row corresponds to the bird called Common Chaffinch (*Fringilla coelebs*). Both the call and song for this bird species are recorded in Riesewohld, Schleswig-Holstein. The x and y axes correspond to time (in seconds) and frequency (in Hertz).

of even finer units called syllables. There are bird species that only produce a single syllable while other species can produce different patterns formed by a few syllables. There are yet other bird species that have even elaborate vocabularies with a complex structure of syllables. These complex syllables can be divided into even more elementary units called elements or notes (Catchpole and Slater, 2003). The songs

[1]North Campus, Faculty of Forest Sciences and Ecology, University of Göttingen.

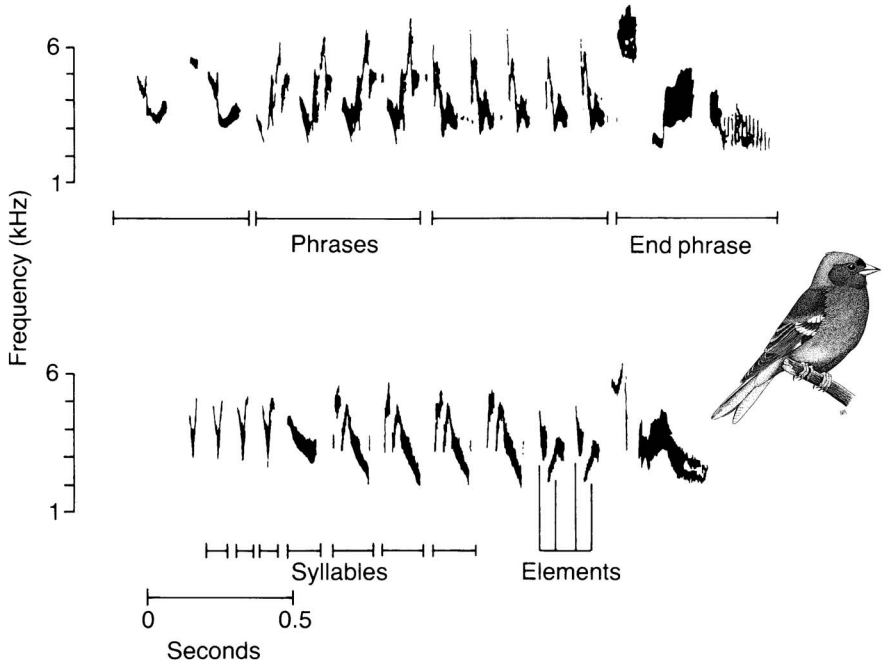


Figure 3.3: Time-frequency representations (spectrograms) of two different songs for the bird species Common Chaffinch (*Fringilla coelebs*). The figure illustrates the hierarchical divisions of bird songs – phrases, syllables and elements. The x axis represents the time in seconds, while the y axis represents frequency in Hertz. The figure has been reproduced from Catchpole and Slater (2003)

produced by the passerines[2] are generally more complex than the songs produced by non-passerines (Potamitis et al., 2014). Fig. 3.3 provides a graphical representation of different sections of a bird song. Different birds vocalize in different frequency ranges and the broad range of most bird vocalizations has been estimated to lie between 250 Hz and 8.3 kHz (Hu and Cardoso, 2009).

3.2 Why are we interested in Bird Vocalizations?

Throughout the timeline of our planet, bird species have been evolving and disappearing due to naturally occurring long-term processes such as earth's climate fluctuations, continental splits, volcanic eruptions, etc. However, it has been reported that the current rate of extinction is outside the scope of natural processes and is heavily driven by environmental changes humans have caused on earth (Lovette and Fitzpatrick, 2016). As per the IUCN Red List of Threatened Species, 1481 bird species are facing an immediate threat of extinction, that is 14% of bird species (iuc, 2021). It was estimated

[2]Birds species that belong to the order Passeriformes are called passerines. More than half of all bird species belong to this order. They are also called perching birds.

in a study conducted across North American avifauna that the net loss of birds in the last 48 years was approaching 3 billion since 1970 (Rosenberg et al., 2019). Therefore, it is becoming increasingly important to monitor bird species trends and population sizes (or for animal species in general) for conservation purposes. We can only take proper measures toward prevention of such extinctions if we know the real status of animal diversity. Birds are relatively well studied because they are found in most environmental habitats. This also makes avian population trends a useful indicator to measure the health of an ecosystem.

The monitoring of bird species diversity has classically been carried out using point counts (or call counts) where a bird expert manually counts birds in a field based on visual and aural cues. This observer-based technique to estimate bird population trends has several limitations ranging from high financial costs to high temporal constraints and errors induced due to reliance on observer expertise (Kahl et al., 2021b). This is where automated monitoring based on bird vocalizations provides a cost and effort-effective way to study changes in avian diversity. Automated acoustic monitoring involves the use of autonomous recording units[3] that can provide continuous streams of real-time data to access the status of bio-diversity in a region with minimal human intervention (Potamitis et al., 2014). These recorders can be installed in fields and left unattended to record bird sounds for weeks or even months. Some comparative studies have shown that autonomous recordings are capable of detecting more species than observer-based surveys (Cunningham et al., 2004; Shonfield and Bayne, 2017; Priyadarshani et al., 2018). Furthermore, the analyses can always be repeated and results reproduced in contrast to observer-based point counts.

3.3 Automatic Bird Detection and Classification

The next challenge after collection of such huge amounts of acoustic data using unattended autonomous recorders is to process the data. Evidently, if this work would have to be done manually, it would be extremely tedious as it would entail going through tens of hundreds of hours of recordings that have been recorded, sometimes for months on end. Pattern recognition of bird vocalizations aims at automating the task of detecting and identifying bird species in field recordings. Therefore, it has become an increasingly common and effective method in the context of bird species monitoring. Notwithstanding the advantages of using bird vocalizations to infer ecologically relevant information, there are certain challenges associated with processing field recordings to produce robust results. Unattended field recordings can be quite

[3]These are self-contained recording devices with high memory capacity and long battery life that are deployed in terrestrial environments for passive acoustic monitoring.

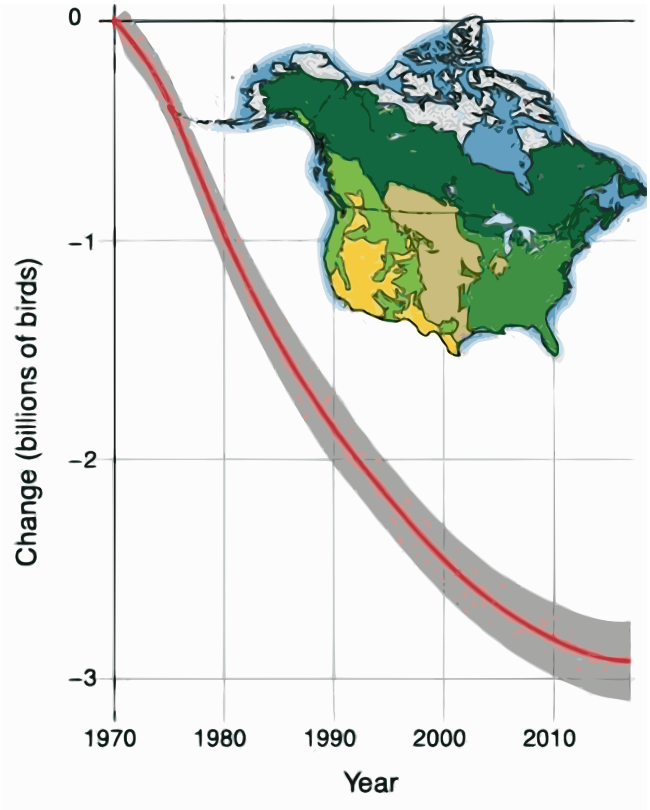


Figure 3.4: Rosenberg et al. (2019) have integrated bird population size estimates and trajectories for 529 species across North American avifauna and reported an estimated net loss of 2.9 billion birds since 1970. Gray shape represents the 95% credible interval. The graphic has been reproduced from the same work.

noisy, as depending on the distance from the recording device, sound samples can be faded or distorted and recordings can include overlapping sounds from the same or different bird species. For this reason, the automated analyses of bird sounds has shown limited success until now. On surveying the literature available on automatic audio-based analysis of bird species, it has been found that several authors have addressed the influence of noise by adding artificial stationary and non-stationary noise to recordings which is unlikely to represent the real world environment. Moreover, most analysis have been performed using less noisy recordings and relatively small datasets, as has already also been pointed out in Priyadarshani et al. (2018). In many works, bird species diversity data has been manually extracted from a reduced set of recordings (Furnas, 2020; Swiston and Mennill, 2009), or the number of species that have been extracted from huge amounts of data has been severely restricted (Wood et al., 2019).

Zhang and Li (2015) have used an *Support Vector Machines* (SVM) classi-

fier trained on Mel-scaled wavelet features where twenty 2s recordings with clean sound samples for each class and added artificial noise are used to classify 30 bird species. The highest average classification accuracy at different SNRs of 85% is reported. Jančovič and Köküer (2011) used a *Gaussian Mixture Model* (GMM) trained on MFCCs/tonal-based features where the recordings were manually split into individual syllable groups. 83% recall performance was reported when white noise with an *signal-to-noise ratio* (SNR) of 10dB was added to the data. Fox et al. (2006) have trained a neural network on MFCCs to report a precision of 89% for 3 bird species. Potamitis et al. (2014) used Hidden Markov Models with perceptual LPCC features to recognize recordings for 2 bird species and reported a precision of about 85%. Somervuo et al. (2006) use sinusoidal modeling and MFCCs and report their best accuracy result of 71.3% on 14 common North-European bird species. Fagerlund (2007) use global decision tree in combination with SVM trained on a set of low level signal parameters and MFCCs reporting the best accuracy of 98% on a set of 8 bird species. Jančovič and Köküer (2015) employed an *Hidden Markov Model* (HMM) in combination with an estimation of frequency tracks to report an accuracy of 78% on 30 bird species. Bastas et al. (2012) proposed a novel Spectrogram-based Image Frequency Statistics feature extraction algorithm and reported an accuracy of 94% for 5 bird species. Lopes et al. (2011) employ a set of classifiers, namely Naive Bayes, the *k-nearest neighbors algorithm* (k-NN), decision tree classifier, *multilayer perceptron* (MLP) neural network, *Sequential Minimal Optimization* (SMO) (Polynomial) and SMO (Pearson). The models are trained on *Music Analysis Retrieval and Synthesis for Audio Signals* (MARSYAS) feature set that includes 64 features in total. The best F_1 score of 65% is reported using an MLP model when full recordings are used for 8 bird species and an F_1 score of 89.7% is reported using SMO (Pearson) when pulses are used to classify the bird species. Damoulas et al. (2010) employed a Bayesian classification algorithm with a dynamic time warping kernel to classify flight calls of 42 bird species. They report an average accuracy of 74%.

Salamon et al. (2017) propose a fusion of shallow and deep learning algorithms to classify different flight calls on the same dataset employed by the previous study (with additional data for same species) and report classification accuracy of 96%. Ruff et al. (2020) employed a deep neural network with four convulational layers followed by two fully connected layers. The model was fed with spectrograms generated from 12s audio recordings for 6 owl species and reported a F0.5 score of 0.5. Sprengel et al. (2016) employ novel data segmentation and data augmentation methods to train a convolutional neural network to classify 50 bird species. They report a mean average precision score of 0.68 when predicting the main bird species in each sound file. LeBien et al. (2020) train an end-to-end convolutional neural network based on mel-spectrograms. The method employs transfer learning to reduce computational

costs, by using a ResNet50 model, pre-trained on the ImageNet dataset. They report a mean average precsion score of 0.893 across 24 species of regional birds and frogs. Cramer et al. (2020) used TaxoNet, a deep neural network architecture with three convolutional layers, a flattening layer, a fully connected hidden layer and a fully connected output layer. They reported a micro-average accuracy of 0.663 for flight calls across 14 bird species. Kahl et al. (2021b) employed a *Residual Neural Network* (ResNet) architecture consisting of 157 layers, 36 of which are weighted. The model was trained using extensive pre-processing and augmentation of sound data. They report a mean average precision of 0.791 for song species recordings across 984 North American and European bird species.

The inception of very deep neural networks has pushed the state-of-the-art higher in terms of classification performance, since hundreds of bird species can now be automatically classified and the methods generalize well. Nevertheless, these networks are not devoid of serious challenges. Firstly, the design of these networks is mostly guided by intuition. Secondly, the training of these networks is tremendously costly in terms of computational resources and time. For instance, the usage of such networks would be an overkill if the training data is limited and there is a limited number of bird species that need to be classified. Furthermore, power-hungry computationally expensive algorithms and power-intensive hardware are not that well suited for real world applications such as mobile recorders (Priyadarshani et al., 2018). In order to realize full potential of passive acoustic data, there is a huge scope to explore computational methods that can provide a better trade-off between the computational costs and overall performance. (Kahl et al., 2021b).

Automated Analysis of Bird Sounds 4

"Simpler solutions are more likely to be correct than complex ones."

William of Occam

4.1 Introduction

In this chapter, different computational experiments aimed at investigating the task of automated analysis of bird vocalizations are discussed; more specifically, the automated classification of bird species based on their sounds. The chapter starts with describing the dataset used. Each experiment begins with description of the experimental setup and analysis tools employed unless they are already explained in Chapter 2. This is followed by results generated and a discussion of the results thereof.

4.2 Dataset

The bird sounds data used in the experiments that will follow in this chapter was provided in the BirdClef 2021 challenge (Kahl et al., 2021a). The dataset was curated from the Xeno-Canto repository of bird vocalizations[1]. Xeno-Canto is a col-

[1] https://xeno-canto.org/

laborative project dedicated to sharing bird vocalizations from all over the world. The dataset consists of short recordings (more than 60,000) of 397 individual birds recorded in South and North America along with metadata with information about location (latitude and longitude), type of vocalization, date, quality of recording, author, etc. The audio data was provided in *ogg* format.

4.3 Computational Experiments

4.3.1 A Simple Statistical Method

In this computational experiment, a simple statistical method was employed to carry out classification. The idea was to benchmark the classification performance without employing conventional machine learning techniques. The classification system consists of two phases: a training phase and a testing phase. Data is prepared by splitting the recordings of varying length into 5 second samples. These audio samples are then resampled to a sampling rate of 22050 Hz. This serves two purposes; firstly, it brings different recordings of varying sampling rates to the same level and secondly, it allows for filtering out frequency parts where the birds do not vocalize. The resampling is followed by peak normalization to adjust the recording based on the highest signal level present in the recordings.

Once the data is pre-processed, it is divided into training and testing subsets. For each bird species, time series of cepstral features are computed (see Chapter 2 for details about cepstral features). To elaborate further, the time series of q^{th} cepstral coefficient $\mathbf{c}_{t,q}$ where $q = 1, 2, 3, \ldots, Q$ is the total number of coefficients for each time frame and $t = 0, 1, 2, \ldots, T$ is the total number of frames in all bird sound recordings in the training dataset for species s. Two different methods are investigated to generate classifications and will be explained as follows.

Comparing Distributions of Cepstral Coefficients

Once the time series of *Cepstral Coefficients* (CCs) are computed, distributions of cepstral coefficients are estimated. For each bird species s, p_i distributions are estimated from the computed time series of cepstral coefficients for individual coefficients q. The idea is to test the assumption that distributions for different species should be representative of the distinct properties of sounds produced by the respective species. Therefore, distributions estimated from recordings sampled from different species should differ more than the ones estimated from recordings of the same species. This should allow us a way to classify a random bird sound recording into one of s class of species. The distributions are estimated by first computing minima and maxima

values of the time series of cepstral coeffiecients. Based on these values, we define $n = 20$ consecutive non-overlapping and equal sized bins to divide the entire range of cepstra values for each coefficient into a series of intervals. The values for different bins are then given by the counts of values that fall into each disjoint bin to provide us an approximate representation of the distribution. In this way, we estimate distributions for the set of species that we want to analyze.

In the testing phase, the same process is followed to estimate the distributions of individual testing samples. This is followed by a simple prediction method that compares the distribution of the test sample with the distributions of individual species estimated using the training dataset. In order to quantify the difference, a common entropy measure is employed called the *Kullback Leibler divergence* (KLD) $D(p_{q,i}||p_{q,j})$ (Kullback and Leibler, 1951), also known as the relative entropy. KLD provides a measure of dissimilarity between two distributions. If $p_{q,i}$ and $p_{q,j}$ are two distributions of a discrete random variable, the KLD is defined as:

$$D\left(p_{q,i}\|p_{q,j}\right) = \sum_{k} \ln\left(\frac{p_{q,i,k}}{p_{q,j,k}}\right) p_{q,i,k}, \tag{4.1}$$

where k refers to the k-th bin of the distribution. $D\left(p_{q,i}\|p_{q,j}\right)$ is defined when both $p_{q,i}$ and $p_{q,j}$ sum up to 1, i.e., both distributions are normalized.

The KLD values computed in this way are summed up:

$$kld_{i,j} = \sum_{q} D\left(p_{q,i}\|p_{q,j}\right), \tag{4.2}$$

to come up with a summarized KLD estimate $kld_{i,j}$ for a test sample i and a species j. The test sample is then assigned the class that corresponds to the lowest summarized KLD estimated $kld_{i,j}$.

Comparing Mean Cepstra

The mean cepstra $\mathbf{C}_q$ are computed by taking the time average for all cepstral coefficients.

In the testing phase, the same process is followed to compute the mean cepstra of individual testing samples. This is followed by a simple prediction method that compares the mean cepstra of the test sample with the mean cepstra of individual species computed for the training dataset. In order to quantify the difference, two simple measures are tested 1) correlation coefficient and 2) Euclidean distance.

Let $\mathbf{C}_q^i$ denote the mean cepstra vector of a species i in the training set that has the highest correlation with the mean cepstra vector of a test sample $\mathbf{C}_q^l$:

$$\rho\left(\mathbf{C}_q^l, \mathbf{C}_q^i\right) \geq \rho\left(\mathbf{C}_q^l, \mathbf{C}_q^k\right), k \neq i, \tag{4.3}$$

where $k = 1, 2, 3, \ldots, N$ and N is the total number of species in the training set. The test sample is then assigned the label i of the species to which the mean spectra vector $\mathbf{C}_q^i$ belongs.

Similarly, let $\mathbf{C}_q^i$ denote the mean cepstra vector of a species i in the training set that has the lowest Euclidean distance to the mean cepstra vector of a test sample $\mathbf{C}_q^l$:

$$d\left(\mathbf{C}_q^l, \mathbf{C}_q^i\right) \leq d\left(\mathbf{C}_q^l, \mathbf{C}_q^k\right), k \neq i, \tag{4.4}$$

where $k = 1, 2, 3, \ldots, N$ and N is the total number of species in the training set. The test sample is then assigned the label i of the species to which the mean spectra vector $\mathbf{C}_q^i$ belongs.

In order to compute the accuracy for the entire test set, we employ the accuracy metric which can be defined as:

$$A := \frac{\sum_{j=1}^{n} c_t(j)}{m \cdot n}, \tag{4.5}$$

where $c_t(j)$ gives the number of correctly classified predictions for a specie j and $m \cdot n$ is the total number of predictions if there are n bird species classes and each class has m test samples.

Results and Discussion

The performance of the classification methods is evaluated by testing for 20 trials drawing 10 randomly selected species, with repetition, for each trial from the larger dataset of 397 species. This is done to arrive at a reliable estimate of performance (for details, please refer to Chapter 5 where this approach is discussed in detail). The results for accuracy of predictions for the three methods discussed above are summarized using box plots in Fig. 4.1. Visual inspection of Fig. 4.1 yields that among the two broad variations of classification techniques used, the one comparing the mean cepstra leads to higher accuracy than comparing distributions estimated over the time series of cepstral coefficients. This is an interesting result since one would expect that, in the latter case, the summarization[2] leads to loss of information which does not happen in the former method since we use all the data to estimate distributions. However, one can argue that averaging over time allows for accentuation of useful trends, components of bird sounds in our case, and similarly leads to diminishing of noisy non-recurring parts in the signals. Besides, since the field recordings are noisy and birds do not vocalize throughout the span of the recordings, using cepstral coefficients computed over all time frames to estimate distributions exposes the distributions to a considerable level of noise. Secondly, since the classifications are

[2] averaging over time

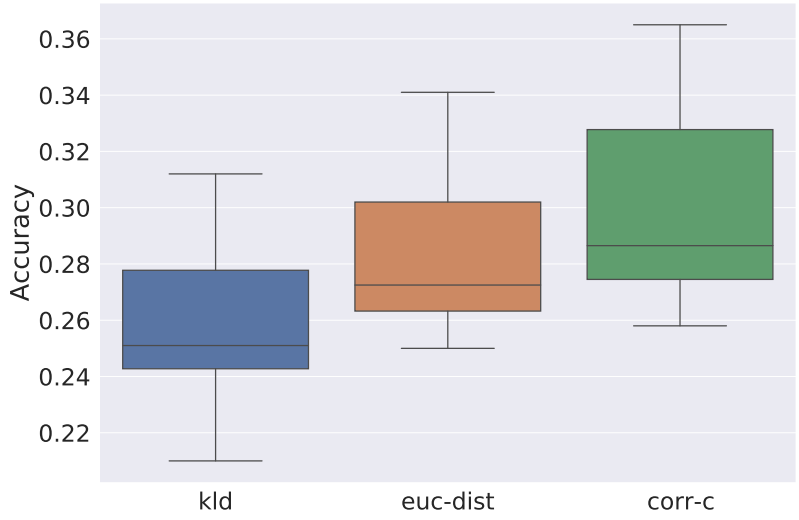


FIGURE 4.1: The accuracy measures for KLD based predictions, Eucledian distance-based method and Correlation coefficient-based predictions estimate. Accuracy is highest when the predictions are obtained by comparing the mean cepstra using the correlation coefficient and worse when distributions estimated from the time series of CCs are compared using KLD. Shown above are the ranges from best result to worst result (whiskers) obtained for 20 different randomly drawn subsets for 10 bird species. The black marking in each box represents the median and the boxes indicate the middle 50% of the results.

generated by comparing the distribution estimated for a test sample with the distributions estimated for the ground truth species where entire training set is used to estimate the distribution, this leads to significant imbalance in data size.

We also observed that using correlation coefficient to compare mean cepstras leads to better performance compared to using Eucleadian distance. One possible explanation could be that a correlation coefficient provides a better estimate of capturing the shape or trend of the data, whereas Eucleadian distance provides an estimate of difference in values of different coefficients. Since cepstral envelope can also provide useful information about bird vocalizations, we achieve a better performance using correlation coefficient.

The overall accuracy of this method is relatively low. Nevertheless the performance is better than a random classifier that will achieve an accuracy of 10% on 10 classes. The performance of this method can be explained by the nature of field recordings that can be noisy, with overlapping sounds of different birds, insects, and other natural phenomena such as wind and rain that interfere with the bird sound signals.

Fig. 4.2 shows distributions for first 6 cepstral coefficients for a set of 10 bird species. Different species are labeled with different colors. One can see from the figure that the distributions estimated from the recordings of different bird species are quite similar which explains why it would not be an easy task for a simple threshold-based classification technique as employed here to classify the sounds of different species.

As observed in Fig. 4.1 one can see that for each method, we get a range of accuracy values, depending on the particular random selection of 10 species. As discussed in Chapter 4, a possible explanation could be the difference in degree of similarity between different subsets of species.

4.3.2 A Random Forest-Based Approach

We learned in the previous section that employing simple statistical techniques might not allow us to model complex patterns in the bird sound data that will allow us to make useful predictions. For this reason, a machine learning-based approach is explored. Compared to conventional methods, machine learning-based approaches have shown an improved performance in a variety of tasks including machine hearing (Bianco et al., 2019).

Following the usual convention, the classification system consists of two phases: a training phase and a testing phase. Both phases comprise of more or less the same steps, namely the data pre-processing, feature extraction, model training, and testing of unobserved data. The data is prepared by splitting the recordings of varying length into 5 second samples. The 5-second audio samples are then resampled to a sampling rate of 22050 Hz. This serves two main purposes: it brings recordings of varying sampling rates to the same sampling rate and filters the parts in the signals where the birds do not vocalize. The resampling is followed by peak normalization to adjust the recording based on highest signal level present in the recordings. Since most birds have been reported to vocalize in the frequency range between 0.5 kHz and 10 kHz (Marler and Slabbekoorn, 2004), we apply a high pass filter to attenuate parts of the signals below 0.5 kHz which strongly reduces the environmental noise. Following the preprocessing of data, a balanced dataset is curated in which 300 sound samples are randomly drawn for each bird species. The data is divided into training and testing subsets wherein 75% data is left for training and the remaining 25% data for testing.

The next step entails extracting low level representations of the data. In addition to the linear scale cepstral features used in the previous experiment, Mel-frequency-cepstral-coefficients (MFCCs) and Mel Spectrogram have also been employed in this experiment. A detailed description of the features can be found in Chapter 2. The features have been calculated for each audio sample using a frame size of 512

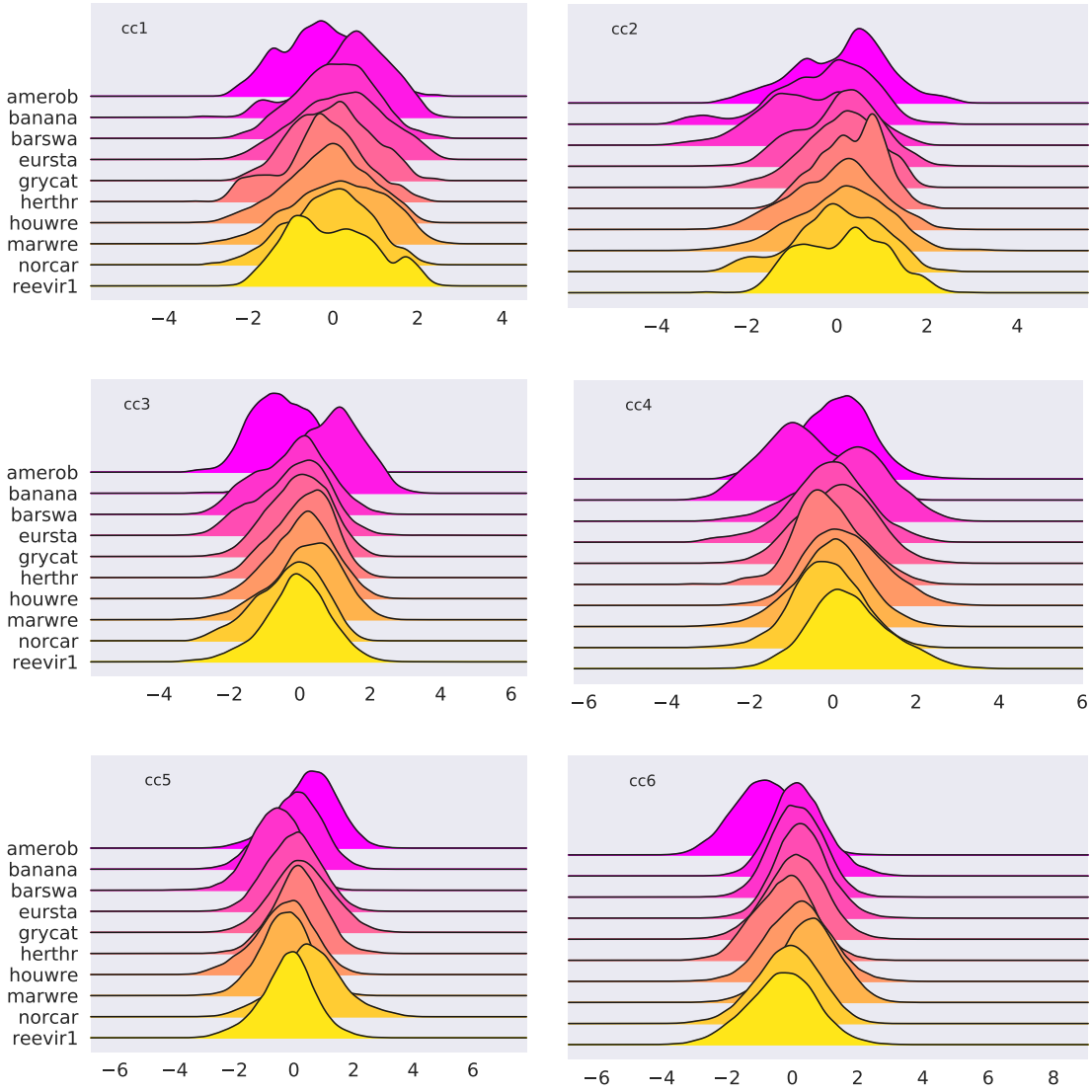


FIGURE 4.2: Distributions estimated from the time series of the first 6 cepstral coefficients $\mathbf{c}_{t,q}$ for a set of 10 bird species, labeled by different colors.

frames with Hamming windowing function while allowing an overlap of 25%. The features thus derived are all time series of feature coefficients. In order to reduce the size of the features so that they can be used as input for the classifier, two different summarization configurations are investigated – summarizing each feature dimension over time by its first two sample moments; mean and variance, and summarizing each feature dimension by its first four sample moments; mean, variance, skewness and kurtosis. The different combinations including the dimensionality of data input to the classifier can be seen in Table 4.1.

To perform the classification, we used a random forest classifier (Breiman, 2001). Our choice for using random forests for the task is inspired by their strong performance in an extensive range of empirical evaluations compared to several other supervised learning algorithms (Caruana and Niculescu-Mizil, 2006). A detailed description of random forests can be found in Chapter 2. For this computational experiment, we have relied upon the random forest implementation provided by the Python library $scikit-learn$ (Pedregosa et al., 2011b). The parameters of the classifier were not tuned manually. While parameters turning can lead to enhanced performance, it also runs the risk of overfitting to the characteristics of the dataset (Stowell and Plumbley, 2014). The classifier was trained with 200 trees and the entropy information gain was used as a quality of split measuring criteria.

Results and Discussion

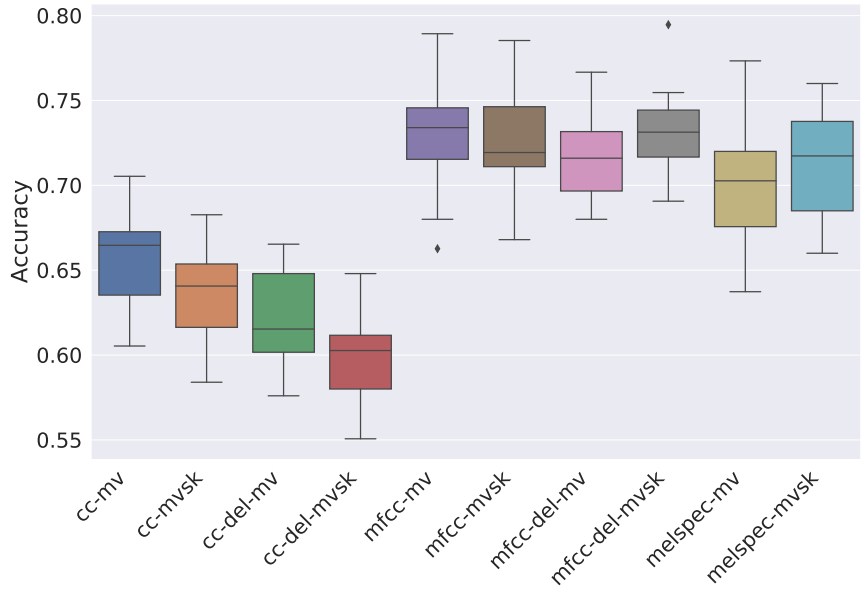
The performance of the classification method is evaluated by testing for 20 trials drawing 10 randomly selected species, with repetition, for each trial from the larger dataset of 397 species. This is done to arrive at a reliable estimate of performance[3]. The model is trained with 10 different feature configurations obtained by varying the type of features used and the summarization employed (Table 4.1). Performance is evaluated using two commonly used metrics, namely accuracy and mean average precision (mAP). The results of different feature configurations are summarized using box plots in Fig. 4.3.

Classification performance is significantly higher than the methods employed in the previous section. Comparing the results obtained using linearly scaled cepstral features, since these features are also used in the previous method, we observed a gain of at least 35% (see Fig. 4.1). This emphasizes the point that these robust machine learning models learn more complex relationships in the feature space to model the decision boundaries compared to what can be achieved by some simple linear comparisons of feature values. The gains are even higher when other two feature types, MFCCs and Mel spectra, are compared.

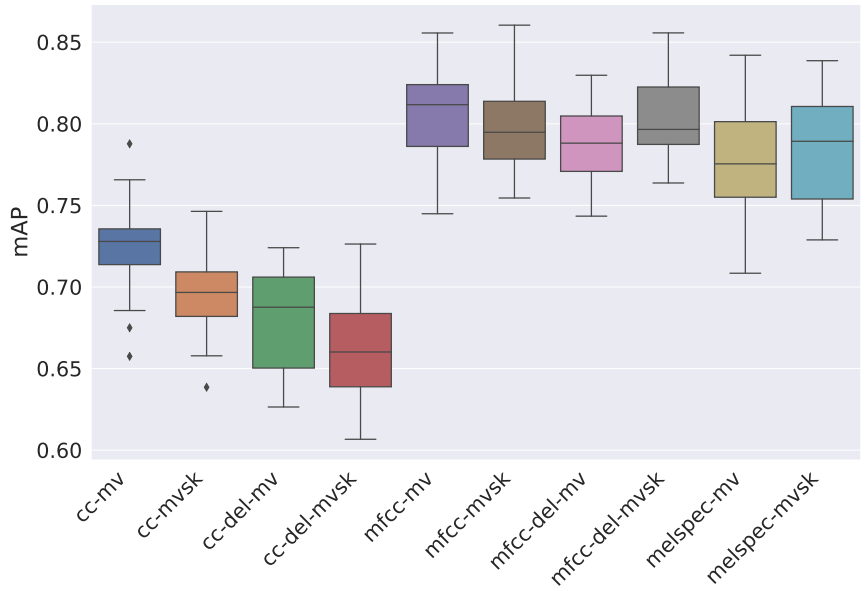
[3]For details, please refer to Chapter 5 where this approach is discussed in detail.

Features	Label	Summarization	Feature Dimension
CCs	cc-mv	mean, var	40
CCs	cc-mvsk	mean, var, skew, kurtosis	80
CCs + deltas	cc-del-mv	mean, var	80
CCs + deltas	cc-del-mvsk	mean, var, skew, kurtosis	160
MFCCs	mfcc-mv	mean, var	40
MFCCs	mfcc-mvsk	mean, var, skew, kurtosis	80
MFCCs + deltas	mfcc-del-mv	mean, var	80
MFCCs + deltas	mfcc-del-mvsk	mean, var, skew, kurtosis	160
Mel spectrogram	melspec-mv	mean, var	256
Mel spectrogram	melspec-mvsk	mean, var, skew, kurtosis	512

Table 4.1: 10 feature configurations are tested, that include three different descriptors, namely, Cepstral coefficients (CCs), Mel-frequency-cepstral-coefficients (MFCCs), and Mel Spectrogram and two feature summarization combinations: mean, variance; and mean, variance, skewness, and kurtosis. The feature dimension is determined by the combination of feature and feature summarization employed.



(a)



(b)

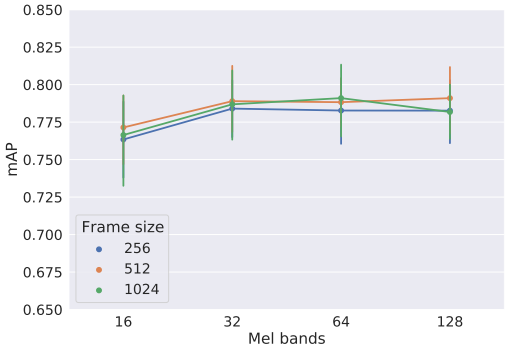
Figure 4.3: Classification results using the performance measures a) Accuracy and b) mAP for different feature configurations using a random forest classifier. Shown are the ranges from best result to worst result (whiskers) obtained for 20 different randomly drawn subsets for 10 bird species. The black marking in each box represents the median and the boxes indicate the middle 50% of the results. Table 4.1 provides the full forms of the abbreviations for the feature configurations used here.

Comparing different feature configurations, it was found that the effect of feature type was the strongest on the classification performance. The linearly spaced Cepstral coefficients were outperformed by both Mel spectras and MFCCs in most trials. The highest mAP value achieved in a trial by a feature type involving cepstral coefficients was around 77%, whereas the highest mAP score achieved by a feature type involving Mel-scaled coefficients was around 86%. Among the two Mel-scaled features, the MFCCs performed slightly better than the Mel spectra, on an average across the 20 trials. This shows that the introduction of mel-scale improves the classification results. In speech recognition, such results are not surprising since the Mel-scale is supposed to mimic human perception, but the fact that better classification results are obtained in case of bird vocalizations is interesting. The MFCCs performing better than the Mel spectras, at least in some subsets of bird species, can perhaps be explained by the fact that the coefficients of Mel spectra are highly correlated. For a classifier such as random forests that employs bagging of different decision trees with different combinations of features, the performance takes a hit if the features are correlated. MFCCs, on the other hand, are decorrelated by first log transforming the coefficients of the Mel spectra followed by a discrete cosine transform.

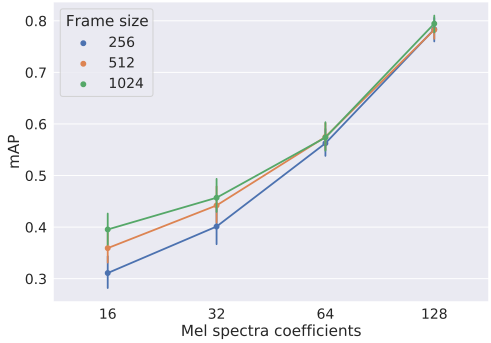
In the case of cepstral coefficients, appending the delta coefficients did not improve the performance. On the contrary, the performance decreased for several subsets of bird species. In case of MFCCs, the classification performance was at best indifferent, in terms of gains, to the effect of appending deltas. The performance did not decrease but didn't increase either.

The effect of the two summarization variations employed in this work was different for the three feature types. For the cepstral coefficients using the third and fourth sample moments, in addition to the first two, led to a dip in performance, at least in some trials. In the case of MFCCs, the results of different trials showed a reverse effect in the case of MFCCs + deltas compared to just MFCCs. In the former case, the median of results from different trials shifted higher by around 2% while the median of results lowered by almost the same degree in case of just MFCCs. In case of Mel spectra, employing the first four moments shifted the median performance over different trials a bit higher. From these results, it can be safely concluded that given the increased computational costs of adding more features, it might be a good trade-off to just use the first two moments for summarizing the time series of features.

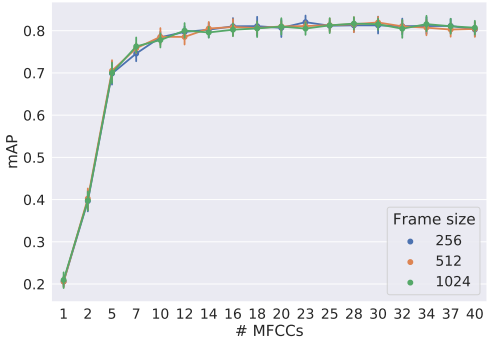
Fig. 4.4 shows scanned parameters for the MFCCs and the Mel spectrogram. The choice of number of coefficients in both Mel spectra and MFCCs had the largest influence on the performance. Using 128 coefficients provided the highest mAP score in case of Mel spectrogram. For the MFCCs, the score increased between 1-18 coefficients and then stayed more or less constant. For the Mel spectra, the variation



(a)



(b)



(c)

Figure 4.4: Classification performance using the performance measure mAP for 10 bird species as a function of different parameters for Mel spectra and MFCCs. a) mAP score as a function of Mel filters and the frame size, when employing Mel spectra. b) mAP score as a function of Mel spectra coefficients and frame size, when employing Mel spectra. c) mAP score as a function of number of MFCCs and frame size, when employing MFCCs. A random forest is used to carry out classification. Shown are the ranges from best result to worst result obtained for 10 different randomly drawn subsets of species.

due to frame size converged as the number of coefficients were increased. Frame size did not influence the performance when employing the MFCCs. The influence of the number of mel filter bands used to compute the Mel spectrogram differs slightly for different frame sizes. The performance increases for all three frame sizes between 16-32 Mel filters. Performance does not change for the frame size 256 beyond 32 filters. For the frame size of 1024, the performance goes up between 32-64 Mel bands and then decreases between 64-128. For the frame size of 512, does not change between 32-64 and thereafter increases slightly between 64-128 mel filters.

4.3.3 Influence of Segmentation

Since the audio recordings of bird vocalizations are usually made in noisy environments and include periods of "silence", or intervals in which vocalizations are absent, it is worthwhile to separate the signal and silent parts of the recordings. We also observed in Section 4.3.1 that when the complete time series was used to estimate distributions without employing any summarization over time, the performance was worse. It was discussed how the presence of environmental noise, especially in periods of silence, can influence the underlying distributions of bird sounds that we were trying to model. Segmentation is a process in which the most representative segments are extracted from the field recordings with the goal to improve the automated identification performance (Colonna et al., 2015; Somervuo et al., 2006).

A segmentation technique developed in parts by Lasseck (2013) and Sprengel et al. (2016) was employed in this work to extract signal parts from noisy audio recordings. Audio signal is first transformed into a spectrogram by applying an STFT. The pixels in the spectrogram that are three times bigger than row median and three times bigger than column median are selected. The intuition behind this is that the high amplitudes in a spectrogram computed from bird sounds usually correspond to bird sound signals. All these selected pixels are set to 1 and all other pixels to 0. Subse-

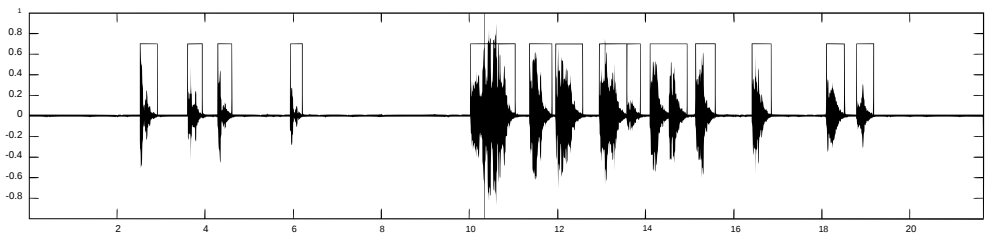


Figure 4.5: Spectrogram of sequence of calls of a Common Kingfisher. Boxes represent the segmented regions, as an ideal result of segmentation. The figure has been reproduced from Potamitis et al. (2014)

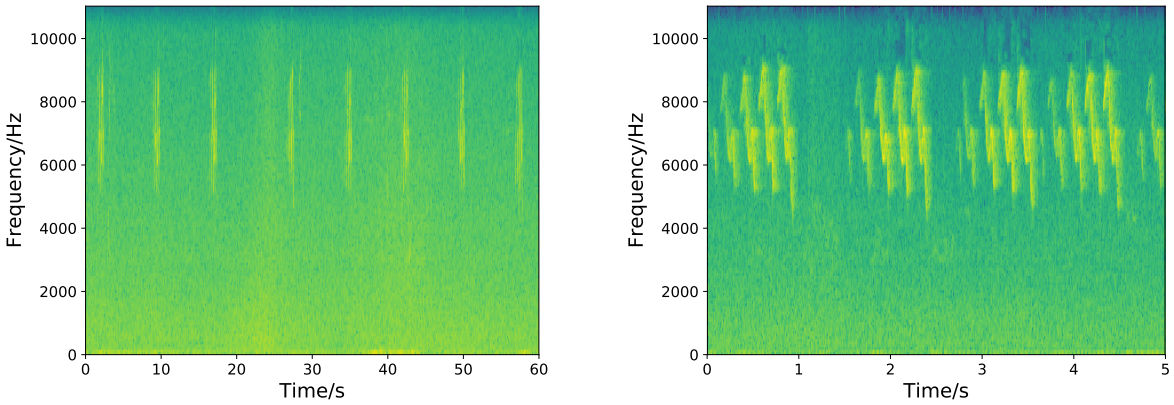


Figure 4.6: Spectrograms of a typical bird song of the bird species American redstart. The left plot corresponds to a 1 minute non-segmented sample. On the right side is the spectrogram of a 5 second segmented sample, where the segments are extracted from the same recording using the segmentation method described above and then the segments are joined to create 5 second samples. The x and y axes correspond to time (in seconds) and frequency (in Hertz).

quently, two mathematical morphology operations are applied to get rid of the noise, namely binary erosion and dilation. 4 by 4 filters are used that are known to produce best results. In order to create a mask that can be used to extract the signal parts from the samples, an indicator vector is created in which the number of elements match the number of columns in the spectrogram. If the i^{th} column in the spectrogram contains at least a single 1, the i^{th} element of the vector is set to 1, else it is set to 0. The indicator function is smoothed by applying two more binary dilation operations (this time a filter size of 4 by 1). The indicator vector is then scaled to the length of the sound file to generate a mask which is then applied to extract the signal parts. These signal parts are simply joined to create a longer signal file, depending on the length we are interested in. In this experiment, the audio recordings are first segmented using the technique described above. Signal segments extracted from the recordings are joined such that 5-second samples are formed. The 5-second audio samples are then resampled to a sampling rate of 22050 Hz. The resampling is followed by peak normalization to adjust the recording based on the highest signal level present in the recordings. Since most birds have been reported to vocalize in the frequency range between 0.5 kHz and 10 kHz (Marler and Slabbekoorn, 2004), a high pass filter is applied to attenuate parts of the signals below 0.5 kHz which strongly reduces the environmental noise. Following the pre-processing of data, a balanced dataset is curated in which 300 sound samples are randomly drawn for each bird species subset. The data is divided into training and testing subsets wherein 75%

data is left for training and the remaining 25% data for testing. Time series of first 20 coefficients of MFCCs are computed. Feature vectors for the samples are then obtained by taking mean and variance of the time series that are then input to a random forest classifier with the same parameters as described in the previous section.

Results and Discussion

The performance of the classification is evaluated by testing for 20 trials drawing 10 randomly selected species, with repetition, for each trial from the larger dataset of 397 species. The model is trained with mean and variance summarization of the first 20 time series of the MFCC coefficients. Features are computed for segmented and non-segmented audio samples to investigate the influence of segmentation. The performance is evaluated using two commonly used metrics, namely, accuracy and mean average precision (mAP). The results are summarized using box plots in Fig. 4.7.

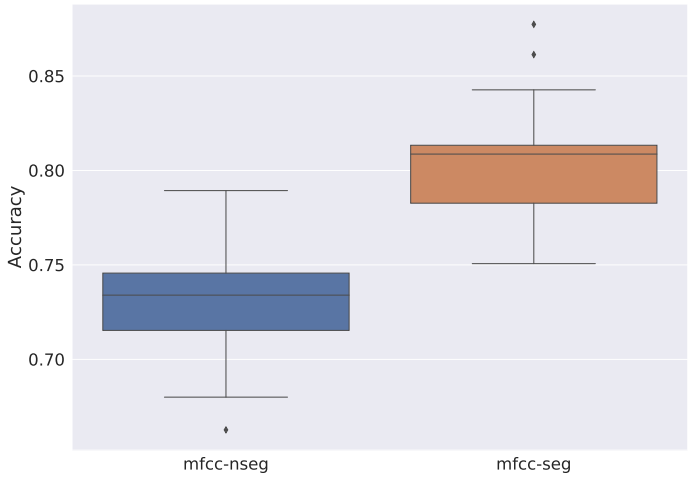
We observe in the figure that the performance increases when segmentation is employed to separate the signal parts from the noise parts of the signal. The highest mAP for a subset in case of non-segmented signals is about 86% and, in case where the audio samples are segmented, the highest Mean Average Precision for a subset within 20 subsets tested is about 93%. This further emphasizes the negative influence of noise on the performance of an automated sound-based analysis system for bird species.

4.3.4 Investigating the Influence of Vocalization Type

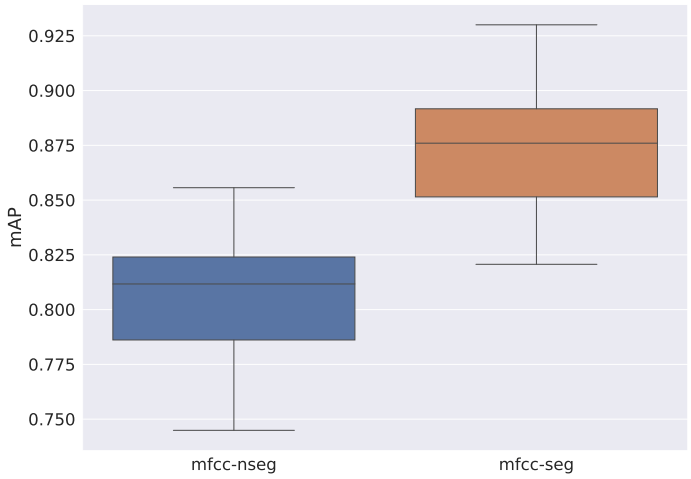
In Chapter 3, we discussed in detail how the bird vocalizations can be divided into two groups – calls and songs. While calls are simpler vocalizations that last for a shorter duration, songs have a more complex structure and are uttered for longer intervals. Calls, therefore, contain less content and are produced in varied circumstances such as alarm calls on sighting a predator, calls signaling information of food sources, etc. Songs, on the other hand, serve a very specific purpose to signal territorial ownership and, more generally, in the context of mating.

Since the structure of calls and songs differs, the influence of these two types of vocalizations on the automated classification performance is investigated in this section. Three separate datasets consisting of all bird vocalizations, only songs, and only calls are curated to train classifiers and the classification results are thus compared.

Audio recordings of the three datasets are first segmented using the technique described in the previous section. The signal segments extracted from the recordings are joined such that 5-second samples are formed. The 5-second audio samples are then resampled to a sampling rate of 22050 Hz. Resampling is followed by peak normalization to adjust the recording based on the highest signal level present in the



(a)



(b)

Figure 4.7: Classification results using the performance measures a) Accuracy and b) mAP with and without applying segmentation to the data using a random forest classifier. Shown are the ranges from best result to worst result (whiskers) obtained for 20 different randomly drawn subsets for 10 bird species. The black marking in each box represents the median and the boxes indicate the middle 50% of the results.

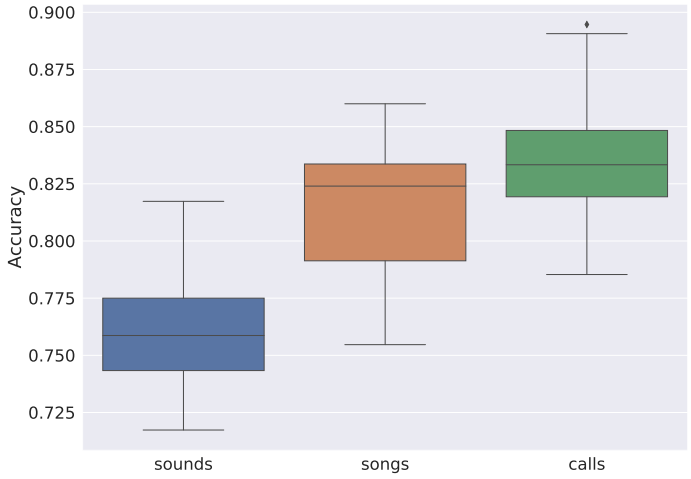
recordings. Since most birds have been reported to vocalize in the frequency range between 0.5 kHz and 10 kHz (Marler and Slabbekoorn, 2004), a high pass filter is applied to attenuate parts of the signals below 0.5 kHz which strongly reduces environmental noise. Following the preprocessing of data, a balanced dataset is curated in which 300 sound samples are randomly drawn for each bird species subset. The data is divided into training and testing subsets with leaving 75% data for training and the remaining 25% data for testing. Time series of first 20 coefficients of MFCCs are computed. Feature vectors for the samples are then obtained by taking mean and variance of the time series that are then input to a random forest classifier with the same parameters as described in the previous sections.

Results and Discussion

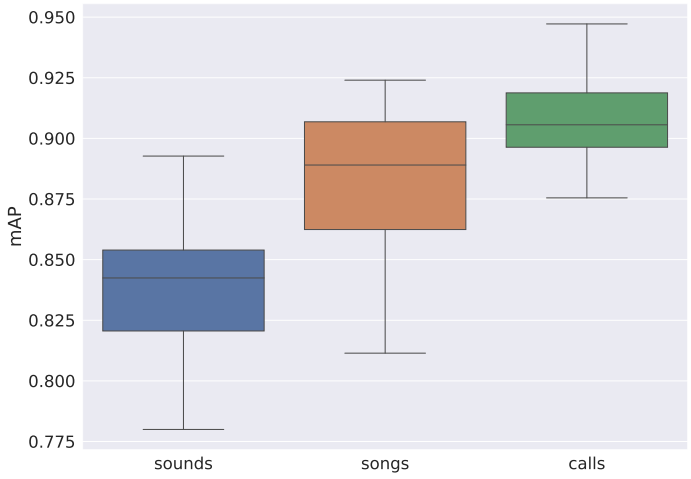
The performance of the classification for three Sets of data is evaluated by testing for 20 trials drawing 10 randomly selected species, with repetition, for each trial from a dataset of 51 species. The 51 species were selected after pruning the larger dataset of 397 species in such a way that each species contained at least 300 audio samples in each of the three experimental categories, namely, songs, calls, and vocalizations with both types of sounds. The model was trained with mean and variance summarization of the first 20 time series of the MFCC coefficients. Features are computed for segmented audio samples. The performance is evaluated using two commonly used metrics, namely, accuracy and mean average precision (mAP). The results are summarized using box plots in Fig. 4.8.

We observe from the figure that the performance of classification increases when the recordings containing two types of vocalizations are separated. The highest mAP score for a set of species in case of all sounds is about 88% while in the case of just songs it is 92.5% and about 94.5% for calls. Note that the same subsets of species were used in all three experiments in order to have a proper comparison. In Table 4.2, one can observe that although the increase in performance varies for different dataSets, the performance nonetheless improves when a single sound type is used for classification. The improvement in performance is as high as 10% in some cases. For instance, for Set 18, mAP for both sound types is 78% and for just calls for the same set of species the value goes up to 91%. These results are not surprising since there are variations in the structure of the two bird sound types as explained above and this should make it harder for a classifier to learn a decision boundary compared to data containing a single sound type and therefore less variation in the properties of data.

Another interesting result that can be observed in the plots in Fig. 4.8 and Table 4.2 is that datasets with calls either perform better than the datasets containing songs or at least perform similarly. One possible reason, among others, that could



(a)



(b)

Figure 4.8: Classification results using the performance measures a) Accuracy and b) mAP for data containing both songs and calls, only songs, and only calls using a random forest classifier. Shown are the ranges from best result to worst result (whiskers) obtained for 20 different randomly drawn subsets for 10 bird species. The black marking in each box represents the median and the boxes indicate the middle 50% of the results.

explain this result is, while calls do not vary within the species, the songs can vary by age, geographical location, and the time of year. Bird species songs can show variation in dialects. Moreover, an individual bird species can have a repertoire of multiple song types.

Sets	Sounds	Songs	Calls
Set 1	0.86	0.91	0.92
Set 2	0.83	0.90	0.90
Set 3	0.82	0.83	0.91
Set 4	0.89	0.89	0.94
Set 5	0.86	0.91	0.88
Set 6	0.84	0.92	0.88
Set 7	0.82	0.86	0.88
Set 8	0.81	0.83	0.90
Set 9	0.86	0.90	0.93
Set 10	0.79	0.81	0.92
Set 11	0.81	0.86	0.90
Set 12	0.84	0.90	0.92
Set 13	0.85	0.89	0.89
Set 14	0.85	0.87	0.91
Set 15	0.84	0.91	0.91
Set 16	0.85	0.89	0.91
Set 17	0.83	0.89	0.88
Set 18	0.78	0.82	0.91
Set 19	0.85	0.91	0.92
Set 20	0.88	0.92	0.95

Table 4.2: Classification results using the mean average precision for data containing both songs and calls, only songs, and only calls using a random forest classifier. Shown are mAP scores for different subsets obtained for 20 different randomly drawn subsets for 10 bird species.

4.3.5 Investigating the Influence of Quality of Recordings

In past bird sounds have been recorded manually, which allows capturing good quality recordings, provided the recordist has sufficient domain knowledge when it comes to recognizing bird species and can handle the recording gear well. This has changed after the advent of autonomous recording units that can be installed in the field and can provide continuous unattended streams of real-time data (Priyadarshani et al., 2016). This is advantageous on several counts when compared to the traditional observer-based method. For instance, the reliance on expert birdists is not required anymore to carry out recordings and birds do not need to be approached in order to record, etc. While this method has proven to provide a better trade-off on several counts, it comes with its own challenges. Unattended field recordings thus made in natural environments lend themselves to a high susceptibility to wide range of noise that can be controlled to some extent when the recordings are done manually. Such noise has been broadly classified into three categories: anthrophony, geophony, and biophony. Anthrophony includes noise induced by humanly engineered machines, for instance, vehicles, aircraft and recording devices, etc. Geophony refers to a variety of naturally occurring sounds that are non-biological, for instance, sound produced by wind, rain, running water, and so on. Finally, one of the most precarious types of noise is biophony which refers to sounds produced by all biological beings other than the ones we are interested in e.g., insects, rodents, and other mammals. In case of targeted recordings, where we are interested in a specific species of birds, even other bird species would be regarded as noise (Farina, 2013).

This experiment is devised to investigate the effect of noise on the classification performance. The dataset used in these experiments allows for extracting the rating of the quality of recordings. These ratings are provided by the recordists at the time of uploading recordings to the database and will, naturally, induce some subjectivity. Given the guidelines of rating provided in the Xeno-Canto database, the data for this experiment has been divided into two classes: recordings that are loud and clear, and recordings that are moderately clear.

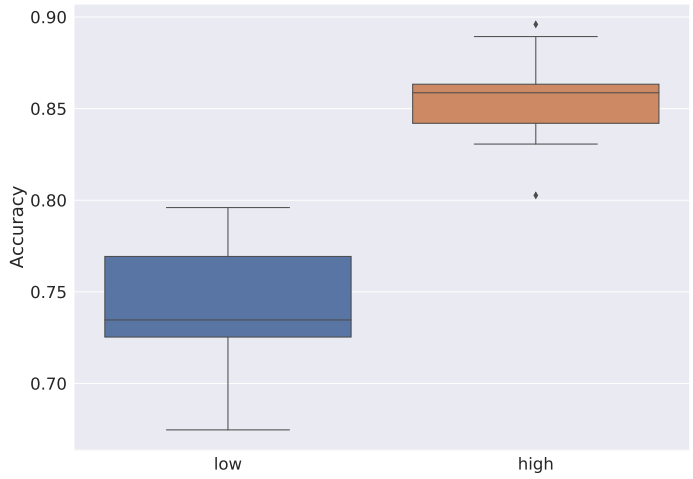
Similar to the previous experiment, the audio recordings are first segmented. The signal segments extracted from the recordings are joined such that 5-second samples are formed. These 5-second audio samples are then resampled to a sampling rate of 22050 Hz. The resampling is followed by peak normalization to adjust the recording based on the highest signal level present in the recordings. Since most birds have been reported to vocalize in the frequency range between 0.5 kHz and 10 kHz (Marler and Slabbekoorn, 2004), a high pass filter is applied to attenuate parts of the signals below 0.5 kHz which strongly reduces the environmental noise. Following the preprocessing of data, a balanced dataset is curated in which 300 sound samples

are randomly drawn for each bird species subset. The data is divided into training and testing subsets leaving 75% data for training and the remaining 25% data for testing. Time series of first 20 coefficients of MFCCs are computed. Feature vectors for the samples are then obtained by taking mean and variance of the time series that are then input to a random forest classifier with the same parameters as described in the previous sections.

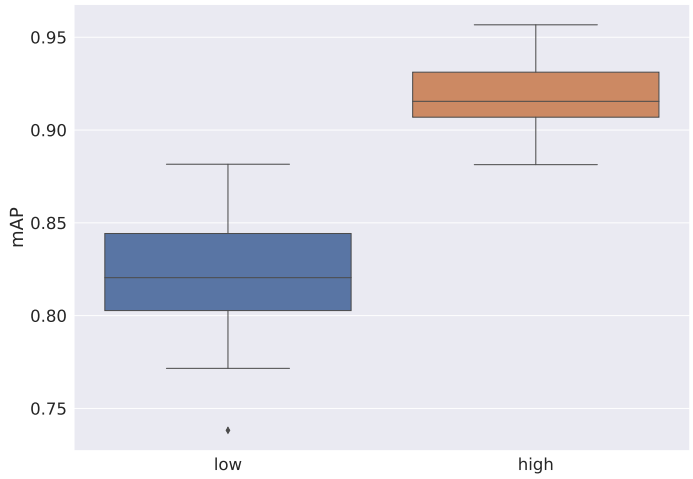
Results and Discussion

The performance of the classification for the two Sets of data is evaluated by testing for 20 trials drawing 10 randomly selected species, with repetition, for each trial from a dataset of 45 species. The 45 species were selected after pruning the larger dataset of 397 species in such a way that each species contained at least 300 audio samples in both experimental categories, namely, bird sounds of low quality and bird sounds of high quality. Features are computed for segmented audio samples. The performance is evaluated using two commonly used metrics, namely, accuracy and mean average precision (mAP). Results are summarized using box plots in Fig. 4.9.

It can be observed in the figure, that on both metrics, the performance increases when higher quality recordings are used to carry out classification. The highest mAP score for a set of species in case of low quality recordings is about 87% while in case of recordings with less noise, the mAP increases upto 96% for a trial. Secondly, the inter-quantile range for the 20 trials in higher quality recordings is higher than the low quality dataset.



(a)



(b)

Figure 4.9: Classification results using the performance measures a) Accuracy and b) mAP for datasets with different levels of noise using a random forest classifier. Shown are the ranges from best result to worst result (whiskers) obtained for 20 different randomly drawn subsets for 10 bird species. The black marking in each box represents the median and the boxes indicate the middle 50% of the results.

Randomized bag-of-birds Approach 5

Large parts of this chapter have been published in the journal
Applied Sciences (Ghani and Hallerberg, 2021).

"Randomization is too
important to be left to chance."

J. D. Petruccelli

5.1 Introduction

One of the main focuses of this chapter is to look into the robustness of classification results and the reliability of accuracy measures, when bird species are selected in a randomized way. Additionally, the influence of the number of selected species (classes) on different measures for multi-class classification success is investigated (see Sec. 5.3.4). For this work, we use the dataset curated by the organizers of the BirdClef 2019 challenge (Kahl et al., 2019). This dataset contains recordings of 659 bird species from South and North America and has originally been drawn from the Xeno-Canto repository for bird sounds (xen, accessed <2020>).

In this contribution, out of 659 available species, n species are randomly drawn with varying n incrementally between 10 and 300. For each n we generate 20 randomly composed lists of species and for each species we choose 200 recordings of comparable length to generate balanced subsets (as will be explained in more detail

in Sec. 6.2.3). The robustness of classification results is then accessed by training a feed-forward neural network on each subset (see Sec. 5.2.3). Our motivations to choose a shallow feed-forward neural network over very deep networks lie mainly in the model simplicity, lower computational costs, and the relatively small amount of data required to train such networks. We wanted to analyze the classification performance using a simple model that can be trained with handcrafted sound features. The performance of classification is accessed using several measures for classification success such as accuracy, precision, recall, area under the receiver-operator-characteristics curve, and mean average precision (see Sec. 2.3.3). As a consequence of these repeated randomized classification experiments, we can hence also provide box-and-whisker plots for each performance metric (see Sec. 5.3). Additionally, it is possible to infer functional relations between the number of classes and classification success (see Sec. 5.3.2). Furthermore, a discussion is included on how confidently the probabilistic classifier classifies different species. In more detail, we discuss whether the predictions made by the model are proportional to the probabilistic confidence the model assigns to the predictions and how this can be used as a measure to evaluate the performance of the classifier(see Sec. 5.3.3).

5.2 Methods

Our audio-based bird classification framework comprises of three modules: data preparation, feature extraction, and model construction.

5.2.1 Bags-of-Birds Approach: Performing Randomized Classification Experiments

In this analysis we use a dataset containing birds sounds of 659 species provided within the BirdClef 2019 challenge (Kahl et al., 2019), originally drawn from the Xeno-Canto repository for bird sounds. The data is prepared, for our analysis, by splitting all sound recordings of varying lengths into 5-second chunks. The audio clips are then resampled to 22050 Hz with Librosa 0.6 audio processing package (McFee et al., 2019). It has been reported that most birds vocalize in the frequency range of 0.5 kHz to 10 kHz (Marler and Slabbekoorn, 2004). The resampling is followed by peak normalisation. To get rid of sound samples that do not contain any bird sounds, a simple signal-to-noise ratio-based estimate is employed. This estimate ensures, with high probability, that 5-second clips that do not contain bird sounds are discarded (Kahl et al., 2018).

Since one of the aims of this work is to estimate robustness of the classification approach, for each classification run of n species, the subsets of species are not care-

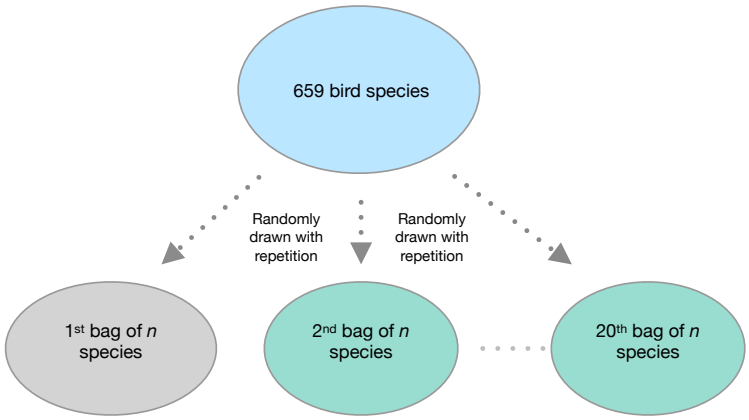


FIGURE 5.1: Schematic for the *bags-of-birds* approach.

fully chosen. Rather, a random number generator is employed to construct 20 bags of species. Analogous to *bag-of-words* approaches[1], one can maybe refer to this procedure as a *bags-of-birds* approach. Each *bags-of-birds* is a set of randomly drawn species without repetition from a complete list of 659 bird species. The idea is to repeat the computation 20 times to have a reliable estimate of classification performance given a certain number of species. Fig. 6.2 illustrates the idea of this numerical experiment. A balanced dataset was curated for the classification task in which 200 sound samples were randomly drawn for each bird species. Finally, the dataset was divided into training and testing sets such that training sets contained 75% and test sets 25% contained of the data. Given that for each bird species we are using 200 sound samples, the test set for each analyzed species will consequently contain $m = 50$ sound samples. This is described in detail in Sec. 2.3.3.

5.2.2 Feature Extraction

The next step entails extracting audio features from time series of audio signals. Extracting features allows us to obtain lower-dimensional compact statistical representations while preserving the distinguishing characteristics of the signal in a non-redundant manner. In addition to reducing the computational costs, feature extraction can maximize the classification performance of the system (Khalid et al., 2014). Studies have shown that aggregated features allow us to achieve a better classification performance compared to a single feature (Xie and Zhu, 2019). In this analysis,

[1] See, for instance Zhang et al. (2010).

we employ the spectral centroid, the spectral rolloff, the zero-crossing-rate, Spectral Bandwidth, the root-mean-square energy, the Mel-frequency-cepstral-coefficients (MFCCs), and MFCCs as features.

Since signal statistics change rapidly, bird sounds, like audio signals in general, are non-stationary signals. For this reason, feature extraction is carried out in a short-term processing manner where the signal is chunked into short *analysis frames*. The analysis frames are assumed to be in a quasi-stationary state (Virtanen et al., 2018). In order to preserve as much information and arrive at a good enough trade-off between frequency and temporal resolution, we select an analysis frame size of 512 samples (23ms) while allowing an overlap of 25%. Therefore, the spectral features described below are computed frame-wise. In case an additional splitting into even shorter windows within each analysis frame was needed (e.g., for the MFCCs), the temporal average of the feature is computed to generate a single value which is associated with the respective analysis frame. For each MFCC, the variance of the coefficients within the analysis frame is computed and used as an additional feature.

To elaborate further, all features are computed using Librosa 0.6 audio processing package (McFee et al., 2019) and can be described as follows (Virtanen et al., 2018): In total the dimension of the feature space is 45, containing first 20 time averaged MFCCs, variances of first 20 MFCCs, time averaged zero-crossing-rate, the spectral rolloff, the spectral centroid, root-mean-square energy, and the spectral bandwidth.

5.2.3 Classification Model

The features are then fed into a feed forward neural network, which is constructed using the sequential model within the Tensor Flow framework (Abadi et al., 2016). Feed forward neural networks are archetypal models for machine learning (Bebis and Georgiopoulos, 1994; Fine, 2006) . In contrast to deep learning approaches used to classify bird sounds, the network consists of only four layers as illustrated in Fig. 5.2. The input to the neural network is the feature vector $\mathbf{x} \in \mathbb{R}^{45}$ which maps through three intermediate layers with $d_1 = 256$, $d_2 = 128$, and $d_3 = 64$ hidden units respectively, and is amplified using *rectified linear units* (ReLU) Goodfellow et al. (2016). Finally, the output layer maps to n independent classes with a softmax transfer function (Goodfellow et al., 2018; Bishop, 2006) where n is the number of bird species.

During training, the model optimizes cross-entropy loss using Adam stochastic optimization algorithm (Kingma and Ba, 2014). We use a constant learning rate of 0.001. To identify the parameter setting that increases the likelihood of predictions, the model is trained for 100 epochs. We observed, that the loss converged to a minimum towards the end of 100 epochs.

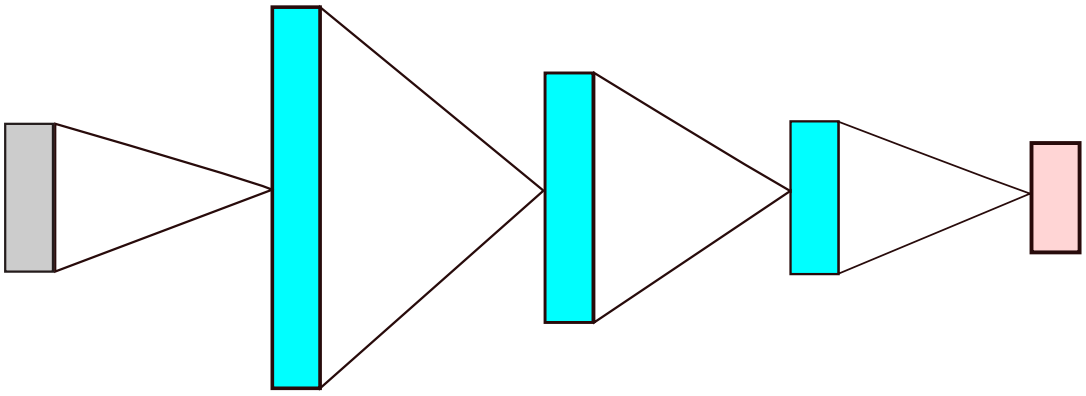


FIGURE 5.2: Architecture of the feed forward neural network used for bird classification. The input $x \in \mathbb{R}^{45}$ maps through three intermediate layers with $d_1 = 256$, $d_2 = 128$, and $d_3 = 64$ and ReLU transfer functions. The output layer maps to n independent classes with a softmax transfer function where n is the number of bird species.

5.3 Results and Discussion

We evaluated the performance of our classification algorithm by testing it for 20 trials on n randomly selected species[2], with n varying between $n = 10$ and $n = 300$. The results for precision, AUC, mAP, recall, and accuracy are summarized in Fig. 5.3. Box plots were estimated from 20 different randomized datasets for each n. Box plots, also known as box-and-whisker plots, provide robust statistical summaries for the data if the sample size is relatively small, i.e., here 20. The box plot divides data into quartiles or fourths – 2 box panels and 2 whiskers. The middle 50% of the data is spanned by the box with 25th percentile or 25% of the data falling below the lower edge of the box (first quartile) and 75% of data falling below the upper edge of box (third quartile). The edges of the box are often referred to as *hinges* and the length of the box is called the *interquartile range* (IQR). The median is indicated by the middle line of box. The whiskers mark the extremes for the remaining 50% of data (Nuzzo, 2016). Surprisingly, we find no increase in the size of the interquartile range for the performance measures with increasing n.

There are several aspects of these results that must be addressed in more detail.

[2]Out of 659 species in the selected dataset.

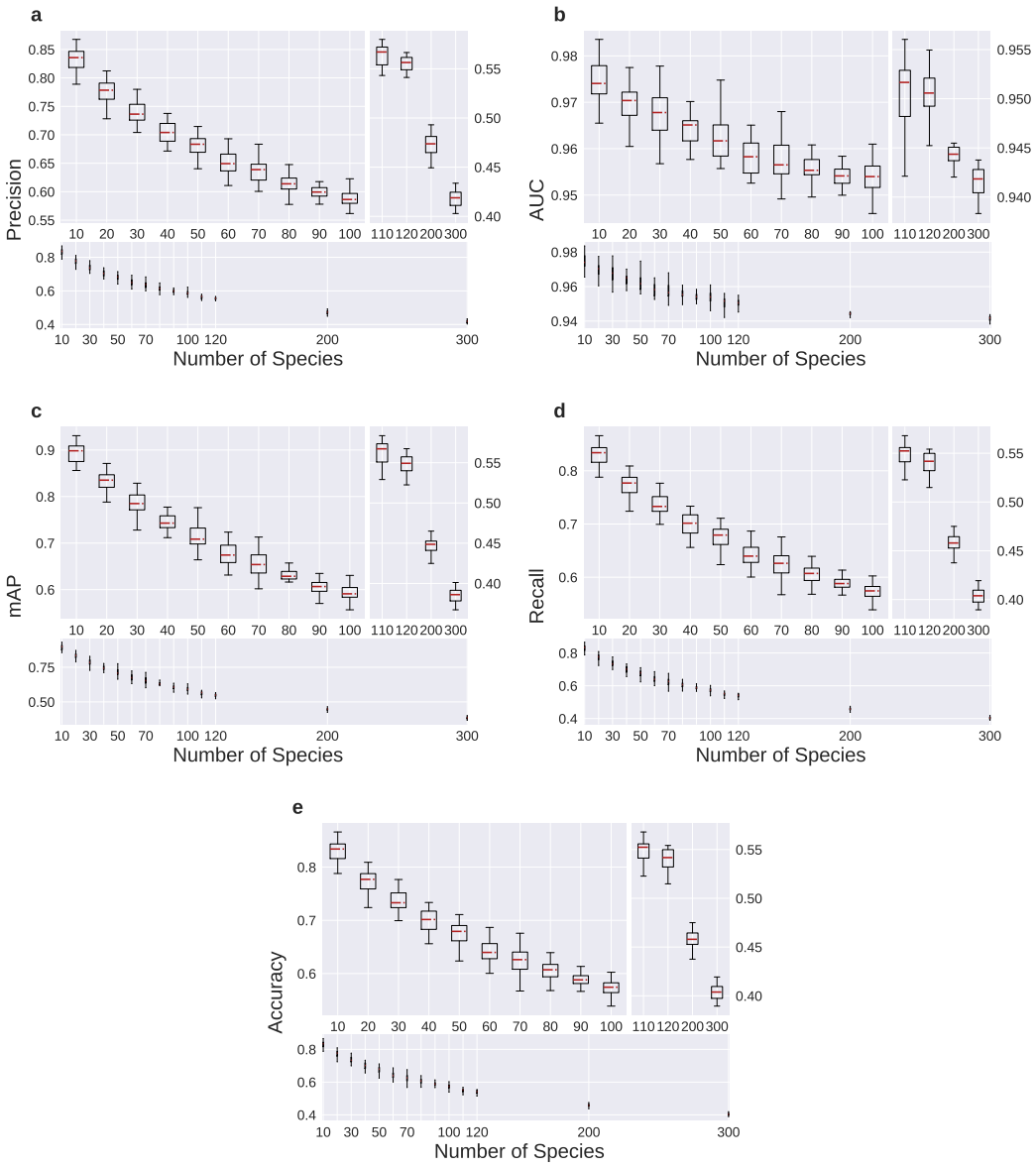


Figure 5.3: The performance measures a) Precision b) AUC c) mAP d) Recall and e) Accuracy decrease as the number of species in each subset is varied between $n = 10$ and $n = 300$. Shown are the ranges from best result to worst result (whiskers) obtained for 20 different randomly drawn subsets for each value of n. The red marking in each box represents the median and the boxes indicate the middle 50% of the results.

5.3.1 Variations Due to Randomized Sub-Sets

As observed in Fig. 5.3, we can see that for each choice of a subset of n species, we get a range of performance values depending on the particular random selection of species. Our results show that the interquartile-ranges vary as much as 12% in some cases. For instance, in case of $n = 30$ in Fig. 5.3(c), we see that the mean average precision (mAP) varies between 0.72 for one subset of 30 species to 0.84 for another subset of randomly drawn 30 species. Similarly, for n = 70, the mAP varies between 0.6 and 0.71. We can see a similar trend in the figures for other metrics considered in this work. As mentioned earlier, the experiment has been repeated 20 times for different randomized selections of n species. The variation in results within different n species' trials shows that classification results can vary significantly depending on the choice of species chosen for analysis. Consequently, it can be inferred that generic claims about the performance of a certain algorithm for a certain number of non-randomly selected species must be interpreted with caution. The results might not generalize for another set of n species even when the species are drawn from the same dataset.

One possible reason, among others, that could explain the variability in performance between different subsets or ensembles of randomly drawn sound samples from n species (*bags-of-birds*) is the possible degrees of similarity of sounds, or the lack of it, between species of different subsets. Ambiguity can be a consequence of similarity of the sounds of species within an ensemble. Therefore, one possible explanation for these results is that sounds of species in the bags leading to lower performance measures have a higher degree of similarity compared to bags that generate higher classification performance.

5.3.2 Dependence on the Number of Species

All performance measures decrease with increased number of species as is visible in Figs. 5.3 and 5.4. An intuitive explanation for this could be that species are more difficult to distinguish when more species are added to the classification task.

However, looking at the definition of the performance measures (Eqs. 2.13 - 2.22), we tried to see if it is possible to understand the numerical results by some analytical reasoning. Consider e.g., the precision $P(n)$ which is defined in Eq. (2.14). If each precision per species $p(j)$ contributing to the average was constant and not depending on n (i.e., $p(j) \sim c$), one should expect $P(n) \sim c$. This is obviously not what is observed in Fig. 5.4. Therefore, one must assume that $p(j)$ is dependent on n, although this is not explicitly visible in Eq. (2.13). To investigate this implicit dependence on

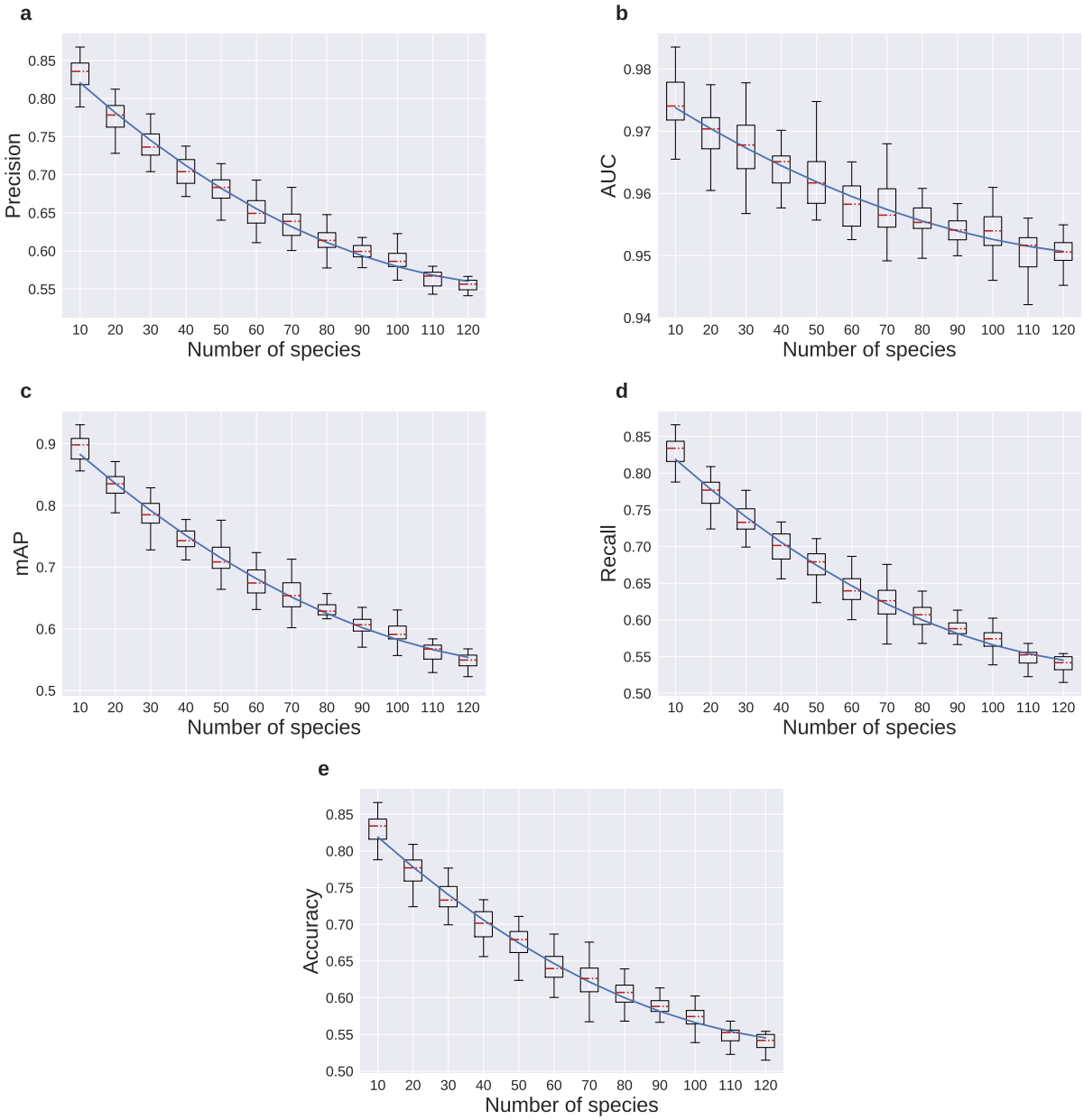


Figure 5.4: The n-dependence of a) Precision, b) AUC, c) mAP, d) Recall and e) Accuracy can be described by fitting quadratic functions (line). The error bars (whiskers) represent the ranges from best result to worst result obtained for 20 different randomly drawn subsets for each value of n. The red marking in each box indicates the median and the boxes show the range of the middle 50% of the results.

n, we have visualized the average numbers of true positives

$$a_{tp} = \frac{1}{n} \sum_{j=1}^{n} c_{tp}(j), \tag{5.1}$$

for each n, and in a similar way, the averaged numbers of false positives, true negatives, and false negatives in Fig. 5.5(a-c). These elements of a confusion matrix enter (in an non-averaged form) into the computed performance measures and thus their dependence on n influences the performance measures. The averaging in Fig. 5.5 was done since the amount of sound samples in each trial depends linearly on n. Therefore, this trivial dependence was removed and we can monitor a non-trivial implicit dependence on n. As is evident, the dependence of the averaged numbers of true positive, false positives, and false negatives can be described relatively well by a quadratic function, whereas the averaged number of true negatives increases linearly with increasing n.

The fact that true negatives, as can be seen in Fig. 5.5(d), behave differently than the other elements of the confusion matrix can be understood by considering the way true negatives are computed in a multi-class classification problem, using a one-vs-all configuration. Each time a sound sample was correctly not classified as the particular species j under consideration, the count of true negatives is increased by one. Therefore, e.g., in a subset of $m \cdot n = 500$ sound samples recorded from $n = 10$ different species, and each species being represented by $m = 50$ sound samples, a perfect algorithm would classify 50 samples correctly as belonging to species j. Consequently, the count of true positives would be $c_{tp}(j) = 50$ and the count of true negatives $c_{tn}(j) = 450$ for a perfect classifier. In other words, we can expect $c_{tn}(j) = nm - m$ with m being the sample size, as specified before, in case of a perfect classifier

$$a_{tn} = \frac{1}{n} \sum_{i=1}^{n} nm - m = \frac{n(nm-m)}{n} = m(n-1). \tag{5.2}$$

The results of the prediction experiments in this contribution with an obviously not perfect classifier reveal that a_{tn} can be fitted by a linear function $a_{tn}(n) = (49.88 \pm 0.82 \cdot 10^{-2})n - (59.27 \pm 0.61)$. Note that the two coefficients are relatively close to the true sample size $m = 50$.

The dependence of the other elements of the confusion matrix on n are more subtle with respect to the range in which these numbers vary and the dependence can be described by quadratic functions

$$a_{tp}(n) = d_2\, n^2 - d_1\, n + c_0, \tag{5.3}$$

$$a_{fp}(n) = -d_2 n^2 + d_1 n + d_0 \quad \text{and}, \tag{5.4}$$

$$a_{fn}(n) = -d_2 n^2 + d_1 n + d_0, \tag{5.5}$$

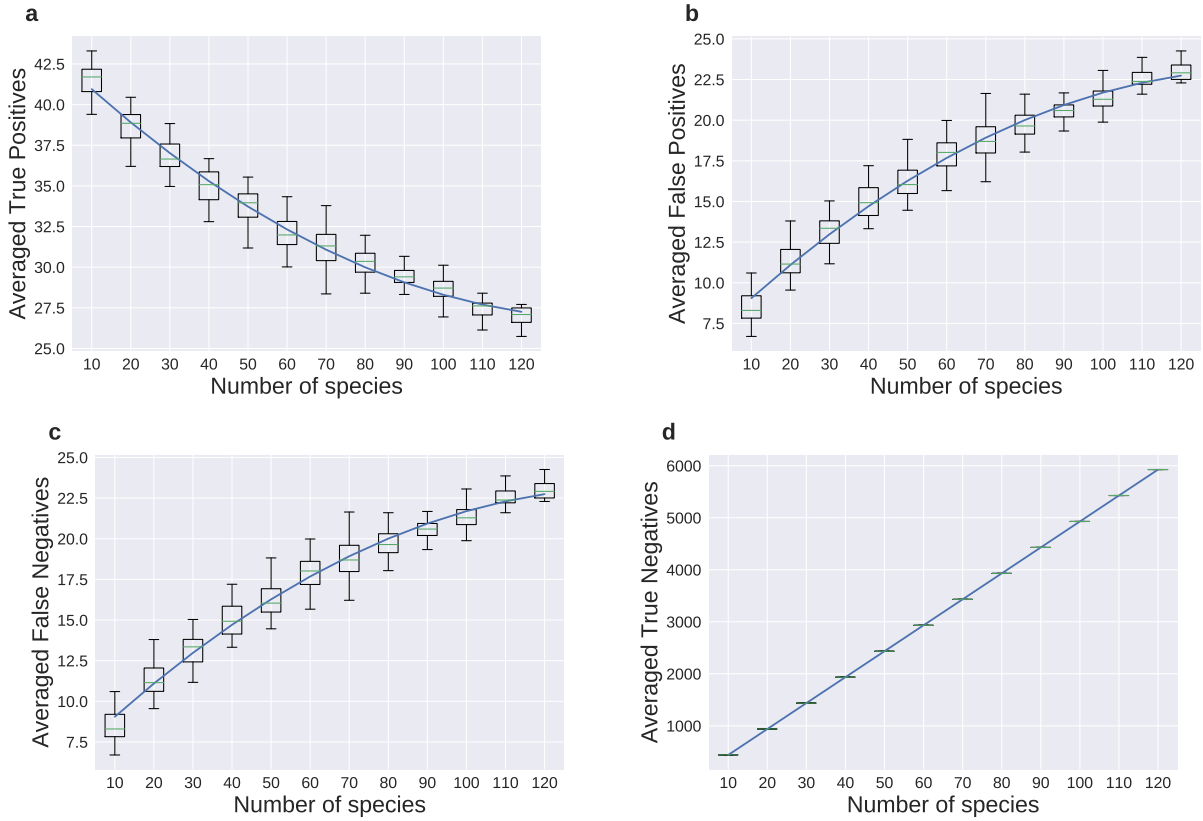


Figure 5.5: The elements of a confusion matrix (a) averaged numbers of true positives, b) averaged numbers of false positives, c) averaged numbers of false negatives, d) averaged numbers of true negatives) display a linear (true negatives) and quadratic (all other elements) dependence on n. The error bars (whiskers) represent the ranges from best result to worst result obtained for 20 different randomly drawn subsets for each value of n. The green marking in each box indicates the median and the boxes show the range of the middle 50% of the results.

with $d_2 = (8.02 \pm 1.00) \cdot 10^{-4}$, $d_1 = (22.87 \pm 1.34) \cdot 10^{-2}$, $d_0 = 6.84 \pm 0.37$ and $c_0 = 43.16 \pm 0.38$. Note that the first two coefficients d_1 and d_2 of a_{tp}, a_{fp} and a_{fn} have either the same values (up to the first 8 digits which are not shown here), or just differ in sign, but not in value. These coefficients are shown in detail here, since we will demonstrate a connection between Eqs. (5.3)-(5.5) and the functions describing the dependence of the overall performance measures.

After describing $a_{tp}(n)$, $a_{fp}(n)$ and $a_{fn}(n)$, one can now try to understand the dependencies of the performance measures. Assuming that each species is classified equally well by a perfect classifier, one would expect $a_{tp}(n) = c_{tp}(j, n)$ for all j and similar for $a_{fp} = c_{fp}(j, n)$ and $a_{fn} = c_{fn}$. Inserting Eq. (5.3) and Eq. (5.4) in

Eq. (2.13) holds

$$p(j,n) \approx \frac{a_{tp}(n)}{a_{tp}(n) + a_{fp}(n)} \approx \frac{d_2 n^2 - d_1 n + c_0}{c_0 + d_0}, \tag{5.6}$$

since the non-constant terms in the denominator cancel each other. Inserting this in the equation for the overall precision (Eq. 2.14) holds

$$P(n) \approx \frac{d_2 n^2 - d_1 n + c_0}{c_0 + d_0} \sim a_{tp}(n), \tag{5.7}$$

since all terms $p(j)$ are identical for the perfect classifier. Consequently, one should be able to predict the scaling of $P(n)$, knowing the coefficients d_2, d_1, d_0 and c_0.

Fitting the coefficients for the quadratic function describing $P(n)$ as in Fig. 5.4, one obtains

$$P(n) \approx g_2 n^2 + g_1 n + g_0, \tag{5.8}$$

with $g_2 = (0.16 \pm 0.02) \cdot 10^{-4}$, $g_1 = -(0.44 \pm 0.03) \cdot 10^{-2}$, $g_0 = 0.86 \pm 0.76 \cdot 10^{-2}$. Note that these coefficients are very close to the coefficients of a_{tp} multiplied with a factor $\frac{1}{d_0+c_0} = \frac{1}{50}$ as indicated by Eq. (5.7). Hence, we could confirm numerically that the dependence of precision on the number of classes follows the dependence of a_{tp} up to a scaling factor of $\frac{1}{d_0+c_0} = \frac{1}{50}$.

Following the same assumptions and reasoning, one obtains

$$R(n) \approx r(j) \approx \frac{d_2 n^2 - d_1 n + c_0}{c_0 + d_0} \sim a_{tp}(n), \tag{5.9}$$

for the recall. Also in this case, the relation between the fitting coefficients of a_{tp} and R is confirmed by the quadratic function fitted to R in Fig. 5.4. Note that for the prediction experiments in this study, the same quadratic function is able to describe the n-dependence of precision and recall.

Extending the above reasoning (i.e., $c_{tp}(j,n) \approx atp(n)$) to explain the n-dependence of the accuracy as given by Eq. (2.17) yields

$$A(n) \approx \frac{1}{m}\left(d_2 n^2 - d_1 n + c_0\right) \sim a_{tp}(n). \tag{5.10}$$

Also, this relation was numerically confirmed by comparing the coefficients for the polynomials describing A and a_{tp}. The values of the coefficients d_2, d_1, and c_0 are given after Eq (5.5).

Discussing the n-dependence of the multi-class AUC and the mAP analytically is not as straightforward as the previous considerations. Therefore only numerical results are presented in this analysis. As one can see in Fig. 5.4, the n-dependence

of AUC and mAP can be also described by quadratic functions. Additionally, we observe that the coefficients for the linear and the quadratic term of the function describing the mAP resemble the coefficients describing $P(n)$ in value[3]. Consequently, one can argue that the above discussion for $P(n)$ could possibly also explain the n-dependence of mAP. Nevertheless, the constant term ($c_0 = 43.16 \pm 0.38$) added to the function describing $P_{mA}(n)$ is higher than the constant offset of the precision.

In summary, we can relate the n-dependence of several measures for the classification success to the n-dependences of the confusion matrix, assuming the behaviour of a perfect classifier we fit functions describing these dependencies. Note that this does not imply that we claim our classifier to be a perfect classifier, neither do we claim that scaling with n which we obtain here is universal in the sense that it will be observed for any other classifier. The latter aspect is a question which needs to be tested in future contributions, but it is out of the scope of this work.

5.3.3 Metric of Confidence

The decisions made by the classifier are based on probabilities which are estimated (through the *Artificial Neural Network* (ANN)) for each species. The predicted label is then assigned to the species with the highest probability. We are aware that a probabilistic classifier does not have to always classify correctly, it just needs to classify with the right frequencies. Following that definition, one way of measuring the performance of the classifier would be to check whether the probabilities it assigns to classifications are really representative of correct classifications. In other words, we need to find out if the frequencies with which the classifier classifies the samples correctly matches with the level of confidence, high or low, the classifier is assigning them. The probabilistic classifier tries to account for implicit uncertainties in the classification process, for instance, due to noisy data, inter-class similarity, inadequate classification capability of the features, etc. Consequently, measuring the performance of the classifier then boils down to measuring the performance of its confidence. That is what we are trying to do here.

We analyze the effect of introducing a *confidence threshold* requiring the assigned probability to be above the threshold in order to accept the classification. In an ideal case, this should follow – if the model assigns a probabilty of 90% to its classifications, then we should see a convergence of 90% for our measuring indices. Apart from measuring the performance of our classifier, this analysis also provides us a way to ascertain where the classifier is over-confident and where it is under-confident. If the probability it assigns is higher than the true positive rate then our model is over-confident and if it gives low probabilities even when the true positive rate is high it

[3]The values describing coefficients of $P(n)$ are given after Eq (5.8).

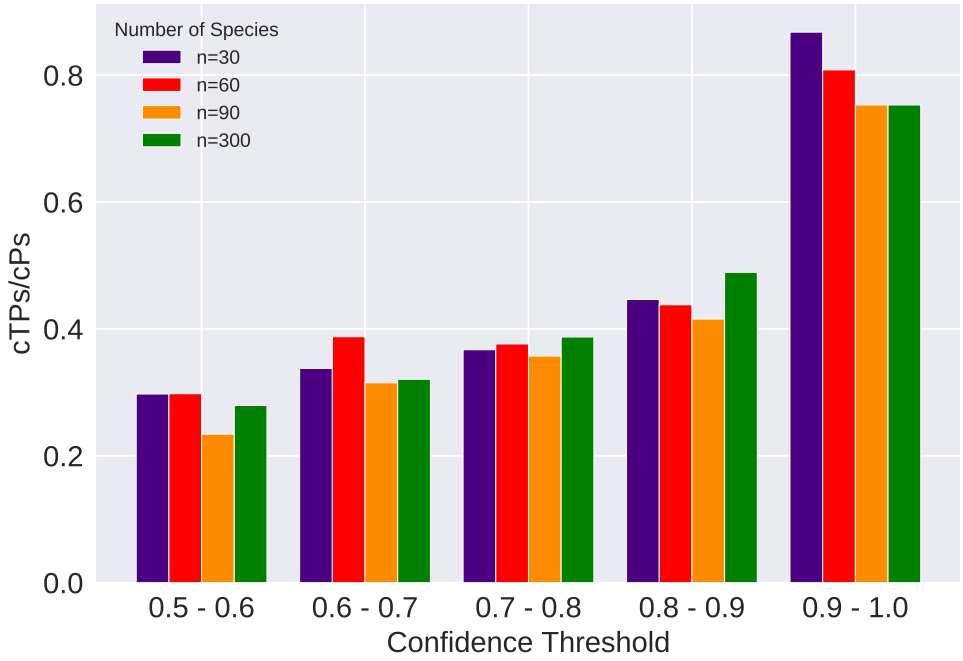


FIGURE 5.6: The precision (cTPs/cPs) increases as the confidence threshold (probability a sample needs to be assigned by the model in order for it to be classified as a positive prediction) is varied between 0.5 and 1.0. Shown are precision values for $n = 30, 60, 90$ and 300. As is evident, for different n, as the confidence threshold tends towards 1.0, precision also increases significantly.

is under-confident. This will allow for targeted improvements in the classification model thereby providing clearer insight on what is actually happening, and provide opportunities for active learning as we can feed the model with more training data of classes that are suffering from over/under confidence. In Fig. 5.6, one can see that the precision (cTPS/cPs)[4], i.e., the ratio of true positives to all classified positives changes as the confidence threshold is varied between 0.5 and 1.0. We observe that precision increases as the confidence threshold is increased. And, for instance, for $n = 30$ species, the precision for confidence threshold in range 0.9-1.0 is more than 0.8. Similar results can be seen for other n. This shows that when the model is assigning high confidence to its predictions, the predictions are mostly correct, which should be expected from a good classifier. Nevertheless, for lower thresholds, it gives high probabilities but is unable to match the classifications. For instance, in the 0.5-

[4]cTPS is the total number of true positives given the confidence threshold and cPs is the total number of positives (true positives + false positives) given the confidence threshold.

0.6 confidence threshold range, we expect the precision to be between 0.5-0.6, but we see a difference of about 0.2 between the confidence threshold and the precision cloud. It gets better towards higher confidence thresholds. The percentage error decreases. This means that the model performs relatively better for samples it is more confident about. And the samples it is not confident about, it errs on the side of over-confidence than being under-confident. Besides, our model assigns high probabilities to most samples classified as positive which basically means most true positives are skewed toward the high confidence threshold. Since, in this analysis, we are defining the measure of the classifier performance by how well it aligns frequencies of true positives with the probability confidence, we can safely claim that our model does not do poorly when it is highly confident about its predictions.

There is work going on in stochastic artificial neural network trained using Bayesian inference instead of traditional backpropagation (Jospin et al., 2020). These models assign probability distributions to weights and have probabilistic activation functions with the intent of getting better insights on the uncertainties associated with underlying processes.

5.3.4 Comparing Different Measures for Classification Success

In this analysis, we use several common measures for evaluating classification performance and compare their results. The primary reason for this is that different indices encapsulate different aspects of the classification performance. Secondly, as mentioned earlier, there seems to be no consensus in the literature available on the choice of evaluation metric for the audio-based bird species classification task. This compelled us to study a set of indices and not rely on a specific metric.

As one can see in Fig. 5.3, the precision and recall for $n < 100$ do not show much disparity but look quite similar. Although, by definition, these two indices encapsulate different aspects of model performance. This can be clearly seen in Fig. 5.7. Here, we see that for different species in one classification run of $n = 10$, the precision and recall values differ. There are species where precision is higher than recall (e.g., species 10) while others where recall is higher than precision (e.g., species 3). But, it seems for $n < 100$, the precision and recall values more or less equalize when an average is taken over species.

From Figs. 5.3 and 5.4, one can additionally see that the accuracy is exactly the same as recall, since the equations of recall and accuracy become the same when an average is taken over all classes.

Additionally, the mean average precision (mAP) was used to evaluate classification success. Increasingly, a number of works in recent years have been using this metric to state the classification performance of their models. Note that average pre-

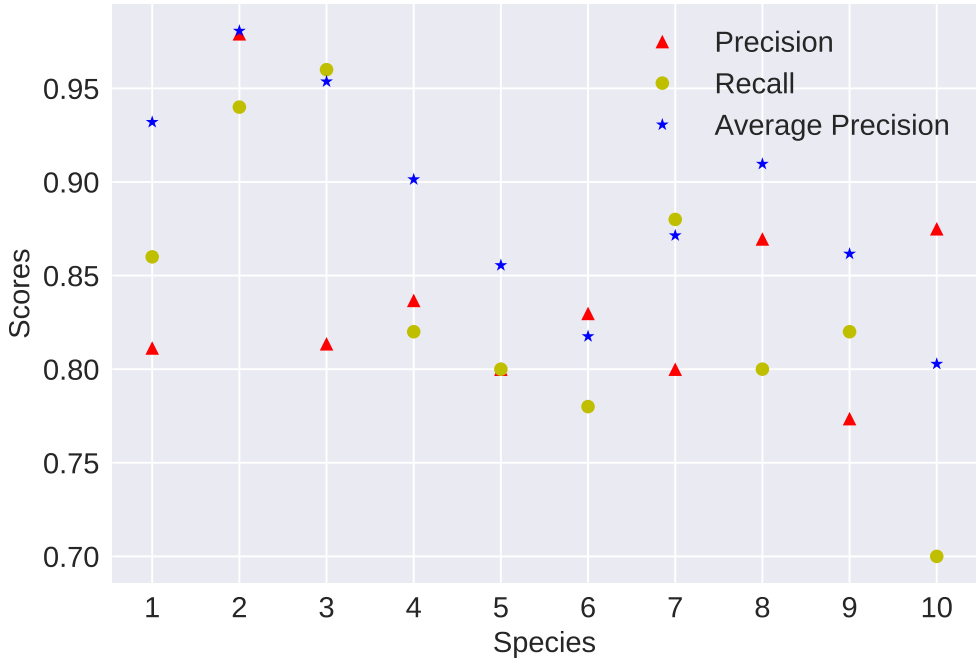


FIGURE 5.7: The precision, the recall, and the average precision for individual species classified in a subset composed of 10 randomly selected species are compared. The largest and the smallest value of each performance measure indicates a relatively large variation and strong dependence on the particular species under study.

cision is one way of measuring the area under the precision-recall curve. Compared to precision and recall, that are computed for one probability threshold, average precision is computed cumulatively by varying the threshold. We see in Fig. 5.3, that although it follows a similar downward quadratic trend as recall and precision, the mAP values are slightly higher than the precision and recall values for different n. For instance, the range for $n = 10$ species for the 20 runs is between 0.86 and 0.94, whereas the precision and recall ranges are between 0.79 and 0.87. This observation is also reflected in the offset of the functions describing the n-dependence as mentioned above.

Another commonly used metric for classification success is the area under the Receiver Operating Characteristics curve (ROC). Our model achieves a high score on the AUC metric as can be seen in Figs. 5.3 and 5.4. Although the AUC score decreases with the increase in number of species n, the score is nevertheless unexpectedly high. For instance, the AUC score for n = 300 for one run is 0.94 which is

unexpected for such a large number of species[5].

In our understanding, the multi-class nature of our problem explains this result. As mentioned earlier, the AUC metric is essentially designed for a binary classifier and has later been generalized for multi-class classification problems (Hand and Till, 2001). Therefore, in case of multi-class problems, one needs to binarize the class labels to compute the AUC score such that the problem is transformed into a binary classification problem with $\frac{n(n-1)}{2}$ binary classifiers (where n is number of classes). Using a one-vs-one configuration (Hand and Till, 2001; Pedregosa et al., 2011a), as recommended by tutorials of many software packages, an AUC score is then computed for each of these binary classifiers and finally an average is computed to get a final AUC score for the entire set of n classes. For an actual binary classifier that classifies poorly, the miss-classifications will reflect in significant enough values of false positives and false negatives to give us a low true positive rate and high false positive rate as per Eq. 2.18 and 2.19. This will result in a low AUC score. However, in the multi-class scenario with one-vs-one configuration, we observe that a classifier distributing misclassifications sparsely across several classes leads to a small number of false positives and small number of false negatives for these artificially assumed binary classifiers. One should note that this will happen even if the classifier fares poorly, i.e., misclassifies with a high rate. An example for this can be seen in Fig. 5.8 which shows a confusion matrix for a classification run with 20 species. It can be seen that the misclassifications are spread throughout the rows and columns of the confusion matrix. Consequently, less numbers of false positives and false negatives will amount to high true positive rate and low false positive rate for individual binary comparisons. Therefore, a high AUC score (refer to Eqs. (2.18) and (2.19)). This is exactly what is reflected on averaging the individual AUC scores to compute the total AUC score for n classes. The classifier is distributing the false predictions sparsely acss several classes and the one-vs-one generalization is unable to capture the actual performance of the model. This leads us to the conclusion that ROC is not a suitable performance measure for multi-class classification tasks. Especially in cases where the misclassifications are distributed rather evenly among several classes, it is very likely to obtain overestimated AUC scores.

5.4 Conclusions

The novelty of this work lies in studying the dependence of classification success on the number of species for bird sound classification. Furthermore, the idea is to

[5]Note that as per the definition of AUC, a random classifier making randomized decisions should give a score of 0.5.

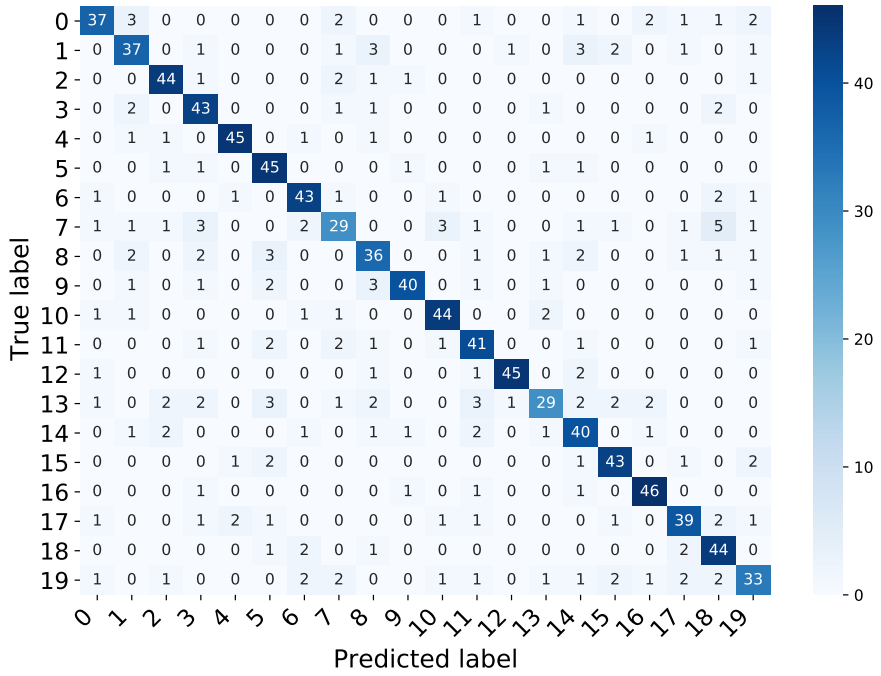


FIGURE 5.8: The confusion matrix for a classification run with $n = 20$ species displays that the misclassifications are spread among several classes. This can result in an overestimation of the averaged AUC for multi-class classifications.

illustrate how these classification results are heavily contingent on the composition of bird species subsets. Therefore, we employ balanced subsets of bird sounds for n species, drawing the species randomly from a larger dataset containing 659 species, where n is varied between 10 and 300. For each n, we repeat the whole procedure[6] 20 times to come up with a reliable estimate of the performance given a certain number of bird species.

The classification is performed using a shallow feed forward neural network trained on 45 pre-computed sound features. We have used a shallow neural network to conduct our analysis primarily due to its model simplicity, less computational costs, and relatively less amount of data that is required to train such networks vis-a-vis deep neural networks. We wanted to benchmark the classification performance and perform our analysis using a simple model that can be trained using hand crafted sound features.

We evaluate the classification performance using several common measures for classification success and also analyze their dependence on n in detail. We ob-

[6]Composition of the subset, training of the classifier and testing.

served that the classification performance is relatively high, even when many different species are present in the datasets under study and using relatively less data. This is an interesting result, since many recent approaches are based on deep neural networks trained on much larger datasets of images of spectrograms without any feature selection. This suggests that shallow neural networks trained on pre-computed sound features can also provide a robust approach to bird classification, which at the same time, is inexpensive in terms of computational costs and the amount of data used.

Concerning the robustness of the approach, we find that all measures of classification success show a decline in value if the number of species present in the subset is increased. For some of these measures, this decline can be explained analytically knowing the n-dependence of the confusion matrix and assuming the behavior of an idealized perfect classifier.

Additionally, we observe that the classification success depends on the individual composition of the bird subsets and classification results can vary significantly depending on the choice of species chosen for the analysis. For this reason, it seems that the generic claims about the performance of a certain algorithm for, say, n non-randomly drawn species, must not be interpreted as a generalized measure of performance for any n species. The classification results might not generalize for another set of n species, even when the species are drawn from the same dataset.

Quantifying Variability in Bird Songs in Widespread Species 6

"There is grandeur in this view of life ... from so simple a beginning, endless forms most beautiful and most wonderful, have been, and are being, evolved."

Charles Darwin

6.1 Introduction

The automatic analysis of bird species based on their vocalizations is an ongoing research topic and several results have been reported towards identification and classification of multitude of bird species around the world. While the newest techniques are improving the state-of-the-art in terms of providing more efficient, robust, and reliable predictions, automated analysis to date has been restricted to within different species. However, there is a growing interest in developing automated methods that can perform separation at an intraspecific level. Such tools can assist studies aimed at understanding variations in vocal patterns within populations of bird species. For instance, the divergence in song characteristics across geographical scales is widespread in several bird species (Catchpole and Slater, 2003), and often leads to regionally dis-

tinct song patterns, many times referred to as dialects (Kroodsma, 2004). It has been proposed that evolution of such variations in vocalizations can arise as a result of adaptation processes of the song features to the conditions of local habitats (Slabbekoorn and Smith, 2002). Another factor that may, over time, result in different song patterns among populations is the social processes associated with the role of learning in the inter-generational transmission of songs (Barros dos Santos et al., 2018; Marler and Tamura, 1962). Such differences have been documented to also occur over very short geographical distances (MacDougall-Shackleton and MacDougall-Shackleton, 2001). Although there is no dearth of studies to investigate this phenomenon, the need for empirical studies to further our understanding of the evolution of variations in bird vocalizations is still there (Petrusková et al., 2015).

Biologists, in order to investigate such phenomena, have been relying on manual approaches to classify vocal variations and dialects, for instance, by visually inspecting the spectrograms of bird sounds to measure spectral characteristics that would allow them to tell differences in vocalizations apart. However, this approach is quite time and labor-intensive and also requires expert knowledge to carry out such analysis. In this analysis, machine learning techniques are employed to classify vocal variation in widespread species in a way that does not require hundreds or thousands of hours of manual processing of recordings.

Two bird species, House Wren and Yellowhammer, have been considered in this work to explore possible variations in songs in the case of former and classification of various established dialects in the case of the latter species. In contrast to existing approaches, these issues are approached in a purely algorithmic manner using the classification approach developed in the previous chapters. To elaborate further, numerous genetic studies have reported that the *aedon*, *brunneicollis*, and *musculus* groups of House Wren have independent evolutionary trajectories (Sosa-López and Mennill, 2014). This information is used as a basis to test the assumption of vocal divergence across these groups in a purely data-driven manner. Furthermore, latitudinal variation in songs of House Wrens is explored by curating 5 regions between $45°$N to $30°$S. In the case of the Yellowhammer species, data available for 7 dialects is used to classify different configurations of these dialects with varying degrees of complexity of classification by choosing. A bag-of-birds approach is employed to randomly create balanced subsets of sound recordings for different classes for repeated classification runs to arrive at reliable estimates of performance. Two machine learning models, a random forest and a shallow feed forward neural network, are employed to carry out classifications. It is observed that the random forest classifier trained on pre-computed sound features is able to classify bird sound variations better than the shallow feed forward ANN. The classification performance is evaluated using accuracy and mean average precision metrics.

6.2 Methods

6.2.1 Species Analyzed

House Wren

The House Wren (*Troglodytes aedon*) is one of the most widely and continuously distributed native bird species in the Western Hemisphere. The latitudinal distribution of the species ranges between Alberta, Canada[1] to Tierra del Fuego[2] (Brewer, 2010). That makes this songbird a unique specimen to investigate variation in vocal patterns across wide geographical scales (Kaluthota et al., 2016). While 30 subspecies have been recognized by the American Ornithologists Union (AOU, 1998), several authorities have grouped the subspecies into five main groups on the basis of geographical and slight morphological differences. Some of these groups have also been treated as separate groups by a few authorities (Howell and Webb, 1995; Navarro-Sigüenza and Peterson, 2004; Kroodsma et al., 2005). The five groups are as follows (Sosa-López and Mennill, 2014): 1) The *aedon* group that includes two subspecies, *T. a. aedon* and *T. a. parkmanii*. The first one is located in southeastern Canada and eastern United States and the second one ranges from southwestern Canada to Baja California, Mexico; 2) The *brunneicollis* group has three subspecies, *T. a. cahooni* that ranges between the mountains of southern Arizona to central Mexico, *T. a. nitidus* found in mountains of Zempoaltepec, Oaxaca, and *T. a. brunneicollis* found in the south of the Sierra Madre del Sur of Oaxaca; (3) The *musculus* group with the most subspecies includes 20 subspecies, covering most areas from the south of central Mexico to Tierra del Fuego; 4) The *beani* group including one subspecies is found in the Cozumel Island; and 5) The *martinicensis* group that includes six subspecies, each of which is restricted to an island in Lesser Antilles. Sosa-López and Mennill (2014) has reported based upon numerous genetic studies that the *aedon*, *brunneicollis*, and *musculus* groups have independent evolutionary trajectories.

There are two distinct sections in a House Wren song. The initial section which is relatively soft is made up of broadband notes that are compared to the terminal section rather unstructured. Notes in this first section are either tonal or harsh, containing multiple harmonics. The terminal section maintains high structure and is considerably louder where the notes are organised as discrete syllable types (Kaluthota et al., 2016).

[1] 58° in the north
[2] 55° in the south

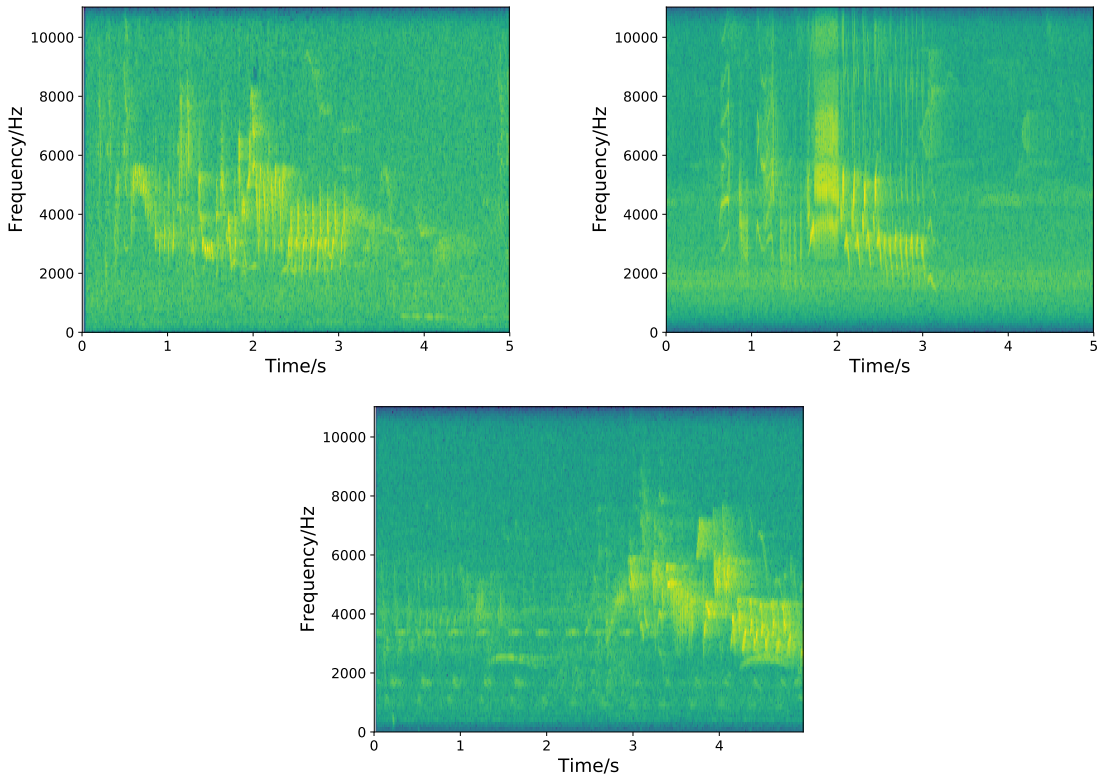


Figure 6.1: Spectrograms of geographically seperated House Wren species. The upper left figure corresponds to a 5-second recording for a typical song of a Northern House Wren (*Troglodytes aedon* group). On the upper right side is the spectrogram for a typical song of a Southern House Wren (*Troglodytes musculus* group). The lower figure is the spectrogram for a song of a House Wren in the *brunneicollis* group. The x and y axes correspond to time (in seconds) and frequency (in Hertz).

Yellowhammer

The Yellowhammer (*Emberiza citrinella*) is a small passerine songbird natively widespread in the Palearctic region, from Central Asia in the east to Spain in the west (del Hoyo, 2001). Nowadays, the bird is typically found in farmlands, and in the Czech Republic, it is one of the most common farmland bird species (Pipek et al., 2018). Yellowhammers have elaborate singing periods, singing from morning through evenings, starting early spring until late summer and even early autumn (Petrusková et al., 2015). A quintessential Yellowhammer song consists of two parts: an initial phrase, which is a quick repetitive sequence of typically 2-21 syllables and the final phrase which is usually composed of two syllables. The terminal part of the song (the second phrase) has been used for dialect distinction since this part has been observed to be shared between males of a local dialect (Pipek et al., 2018). It

FIGURE 6.2: Range map[3] for bird species *House Wren* (Fink et al., 2020)

has been observed that the terminal part of the song is often left unfinished or even not sung at all (Gruber and Nagle, 2010).

Most studies on the dialects of the Yellowhammer have adopted the classification provided by Hansen (1985) based on his study of Danish Yellowhammer populations. In his study, he has characterized different dialects based upon variations of elements in the final phrase (see Fig. 6.4 for spectrograms of different dialects). The different elements have been labeled as B, C, D, E differing in frequency, duration, and modulation. Furthermore, Xl (l for long) and Xs (s for small) vary in duration where the former is a very short sound and the latter long. Bh and Bl indicate relative frequencies of adjacent B elements where h corresponds to a higher frequency and l corresponds to a lower frequency. The different combinations of the these elements, mostly in pairs, form different end-part variants that, as complete songs, are then classified as distinct dialects. Some examples of spectrograms of the Yellowhammer

[3] `https://birdsoftheworld.org`

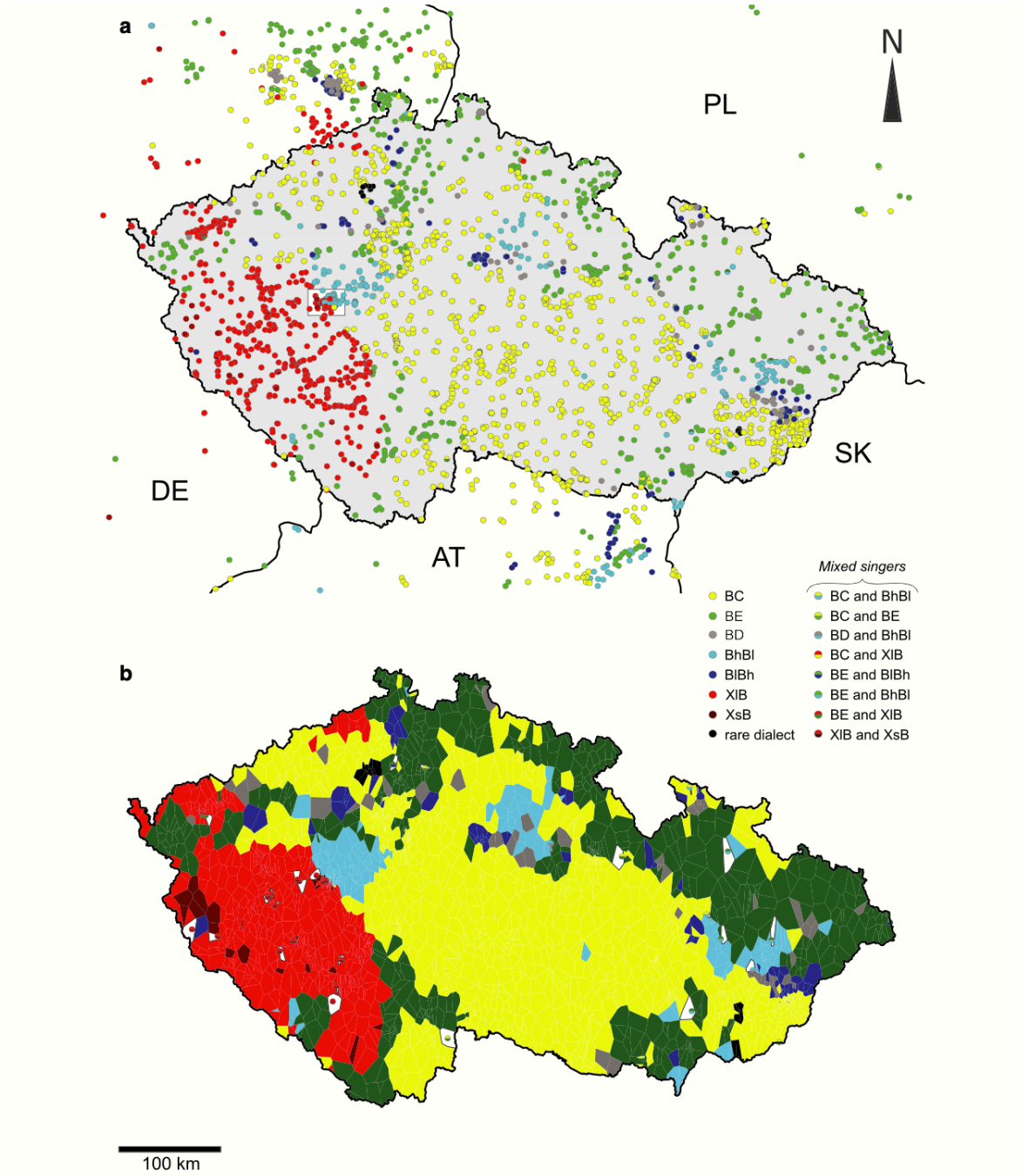


Figure 6.3: Distribution of Yellowhammer dialects in Czech Republic and some data points from neighboring countries. (a) Data points in different colors refer to distributions of different dialects. b) Voronoi polygons are created around map points. Repertoire of mixed singers is indicated by respective colours in white polygons. AT, DE, PL, and SK are labels for neighboring countries Austria, Germany, Poland, and Slovakia, respectively. The figure is taken from Diblíková et al. (2019).

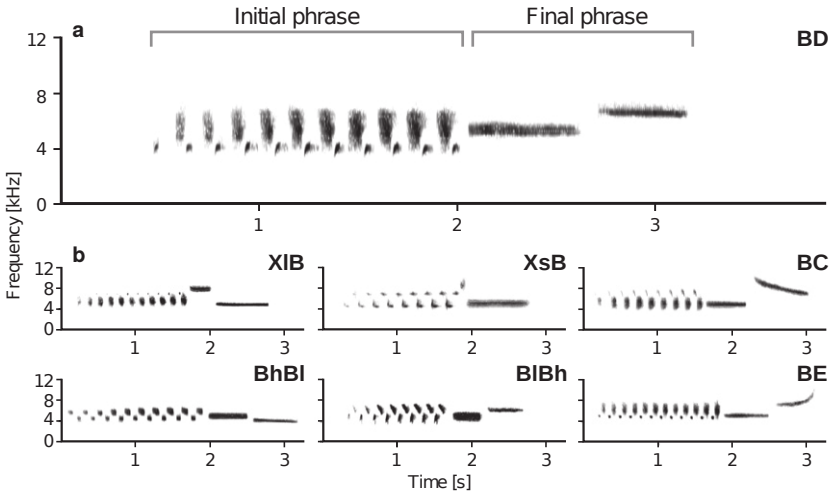


FIGURE 6.4: Spectrograms representation of different *Yellowhammer* dialects. Figure is reproduced from Diblíková et al. (2019).

dialects that are commonly observed in the Czech Republic are BD, BC, BC, BE, $BhBl$, $BlBh$, XlB, and XsB (Diblíková et al., 2019).

6.2.2 Dataset

The songs for the House Wren species have been drawn from the Xeno-Canto repository[4]. The subsets for different groups considered for the classification experiments were curated using the geo-spatial coordinates provided by the recordists for the place where the recordings were made.

For the Yellowhammer species, the dataset was obtained from the researchers working on the Yellowhammer dialects at the Department of Ecology, Charles University, Prague[5]. The dataset included recordings for 7 dialects of the species that were recorded in the Czech Republic. In total, 450 recordings were provided where the time of each recording ranged between a few seconds to a few minutes.

6.2.3 Data Pre-Processing

The data is prepared by first segmenting the audio signals to extract the signal parts from the noisy audio recordings using the technique developed by Lasseck (2013) and Sprengel et al. (2016) and described in Section 4.5. The signal segments extracted from the recordings are joined such that 5-second samples are formed. The

[4] `https://xeno-canto.org/`
[5] `http://www.yellowhammers.net/`

audio clips are then resampled to 22050 Hz with Librosa 0.6 audio processing package McFee et al. (2019). It has been reported that most birds vocalize in the frequency range of 0.5 kHz to 10 kHz (Marler and Slabbekoorn, 2004). The resampling is followed by peak normalization. To get rid of sound samples that do not contain any bird sounds, a simple signal-to-noise ratio-based estimate is employed. This estimate ensures with high probability that 5-second clips that do not contain bird sounds are discarded (Kahl et al., 2018).

6.2.4 Acoustic Analysis

Once the recordings are pre-processed, the next step is extraction of descriptors that provide lower-dimensional compact statistical representations of the bird sound recordings. Time series of first 20 coefficients of Mel frequency cepstral coefficients (MFCCs) are extracted from the audio signals. The feature extraction is carried out in a short-term processing manner where the signal is chunked into short *analysis frames*. In order to preserve as much information and arrive at a good enough trade-off between frequency and temporal resolution, we select an analysis frame size of 512 samples (23ms) while allowing an overlap of 25%. Therefore, the spectral features described below are computed frame-wise. The temporal average of the feature coefficients is computed to generate a single value which is associated with the respective analysis frame. For each MFCC, the variance of the coefficients within the analysis frame is also computed and used as an additional feature. The MFCC features were computed using Librosa 0.6 audio processing package McFee et al. (2019). In total, the dimension of the feature vector is 40, containing first 20 time averaged MFCCs, and the variances of first 20 MFCCs.

House Wren

In this study, we only use *song* vocalizations. Since the aim of this work is to quantify variability in vocalizations for a widespread species, we split the data into different groups. It has been reported based upon numerous genetic studies that the *aedon*, *brunneicollis*, and *musculus* groups have independent evolutionary trajectories (Sosa-López and Mennill, 2014). We use this information as our basis to test the assumption of vocal divergence across these groups in a purely data-driven manner. Consequently, we create three subsets of sound recordings, one for each of the three groups. For the second experiment, we investigate the geographical divergence in vocalization characteristics by simply controlling for latitude. Five geo-spatial regions are demarcated between $45°$N and $30°$S, each region spanning 15 degrees.

We need to ensure that the classifier is learning just from homologous vocalizations and not some other sytemic differences between different geospatial locations.

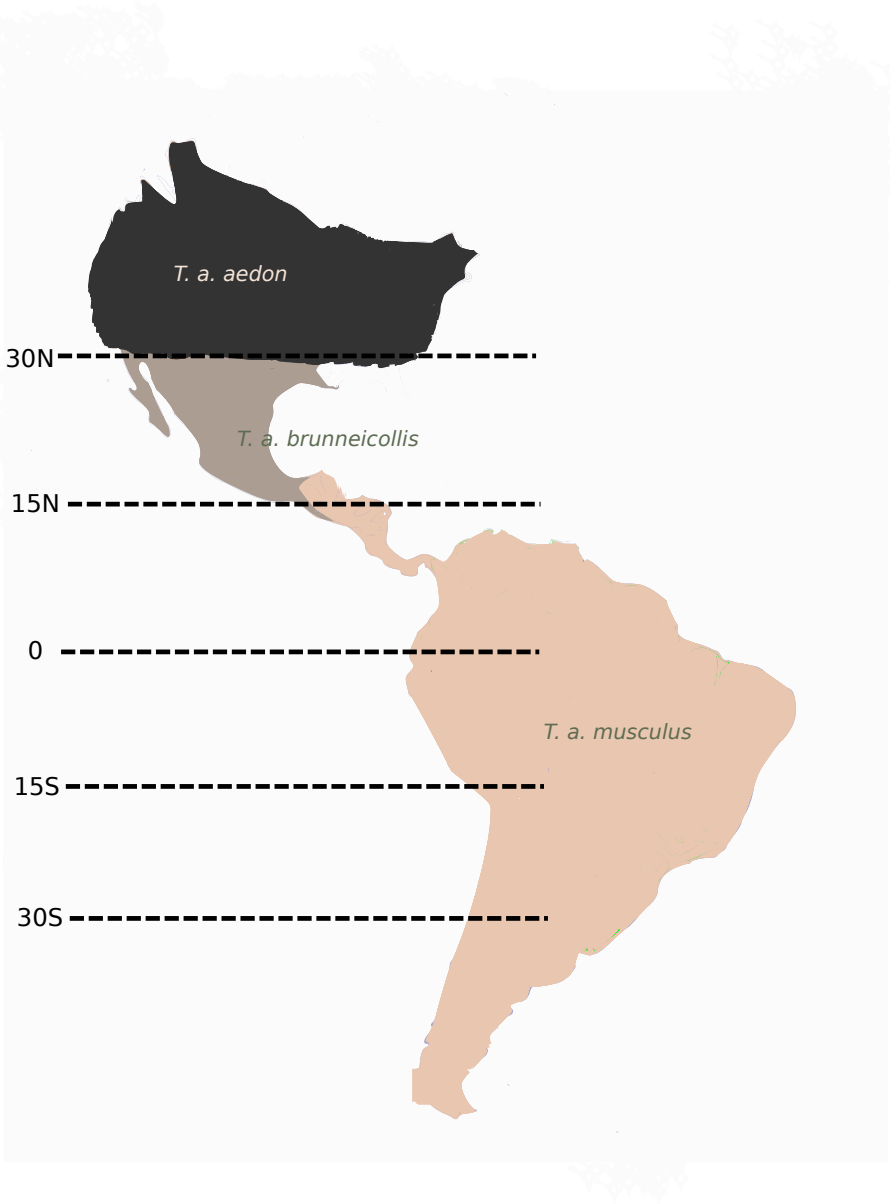


FIGURE 6.5: Map of North and South America, showing the distributions of different groups of House Wren species considered in this study. The different colors in the map refer to areas inhabited by the three groups, namely the *aedon* group (Northern House Wren), the *brunneicollis* group, and the *musculus* group (Southern House Wren), respectively. The dotted lines show the demarcation for the classification carried out while controlling for the latitude.

One factor could be the differences in acoustic properties of typical geo-spatial areas, but since the recordings have been made at multiple sites by different recordists within each area, we can safely assume that the acoustic properties such as background noise, nature and quality of the recording gear, other species present, typical wind conditions, differences in sound attenuation, etc. also vary within an area. Table 6.1 shows the distribution of recordings at different sites in one geospatial area for the bird species *House Wren*. For instance, only one bird sound recording has been made at each of the 89 different sites. Similarly, 2 recordings have been made at each of the 15 different sides and so on. Furthermore, since we are randomly drawing sound clips for training our classifier, this will ensure that we end up with a mixed bag of recordings from different sites with different acoustic properties.

We have relied on the annotations provided by the recordists to determine the vocalization type. The verification of actual vocalization type is out of the scope of this work.

Number of sites	**Number of recordings (at each site)**
89	1
15	2
9	3
3	4
2	5
1	8
1	9

Table 6.1: The data corresponds to recordings made in one geospatial area for the bird species *House Wren*.

Yellowhammer

In the case of Yellowhammer, since the ground truth for different dialects for recordings made in the Czech Republic is available, we use this data to carry out different dialects-based classification. In total, there are 7 dialect types that can be broadly divided into B dialects and X dialects. First, classification is carried out at this level of complexity where data is sorted into two groups based on their labels – all dialects starting with a B element, as a terminal phrase, make the B subset and ones starting with an X element form the X subset. Consequently, the B group consists of BD, BC, BE, $BhBl$, and $BlBh$ dialects and the X group contains the XlB and XsB

dialects. A binary classification is carried out to determine one of the two group types for a sample. The second classification configuration comprises of training the classifier to classify different dialects within these two groups. Therefore, 5 sets of data are formed for B group, while 2 sets are formed for X group. Finally, as a final configuration, the classifier is trained to classify all 7 dialects.

In order to ensure robustness of classification, for each classification run, the subsets of sounds for classification were not purposefully chosen. Rather, a random number generator is employed to construct 20 bags of m sound clips for each group for different classification experiments (see Fig. 6.6). The m is varied for different classification experiments depending on how much data was available for each group. Table 6.2 provides a list of m for different experiments for both species.

Each *bags-of-birds* is basically a set of randomly drawn sounds clips without repetition from a complete set of sounds for each location. The idea is to repeat the computation 20 times to have a reliable estimate of classification performance. Fig. 6.2 illustrates the idea of this numerical experiment. Finally, the dataset was divided into training and testing sets such that training sets contain 75% and test sets contain 25% of the data.

Two machine learning algorithms are employed to perform the classification – the random forest and the feed forward artificial neural network. The parameters in both models were not tuned using a grid search to get the best performance since this often leads to over-fitting. Nevertheless, some parameters were chosen based upon hand

Species	**Classification configuration**	**Number of sound clips per class**
House Wren	Group based (3 classes)	300
House Wren	Group based (2 classes)	300
House Wren	Latitude controlled (5 classes)	300
Yellowhammer	X, B groups (2 classes)	600
Yellowhammer	All dialects (7 classes)	500
Yellowhammer	X dialects (2 classes)	500
Yellowhammer	B dialects (5 classes)	500

Table 6.2: Number of sound clips per class used for different classification configurations tested for House Wren and Yellowhammer. Different classes pertain to different classification configurations employed in different experiments.

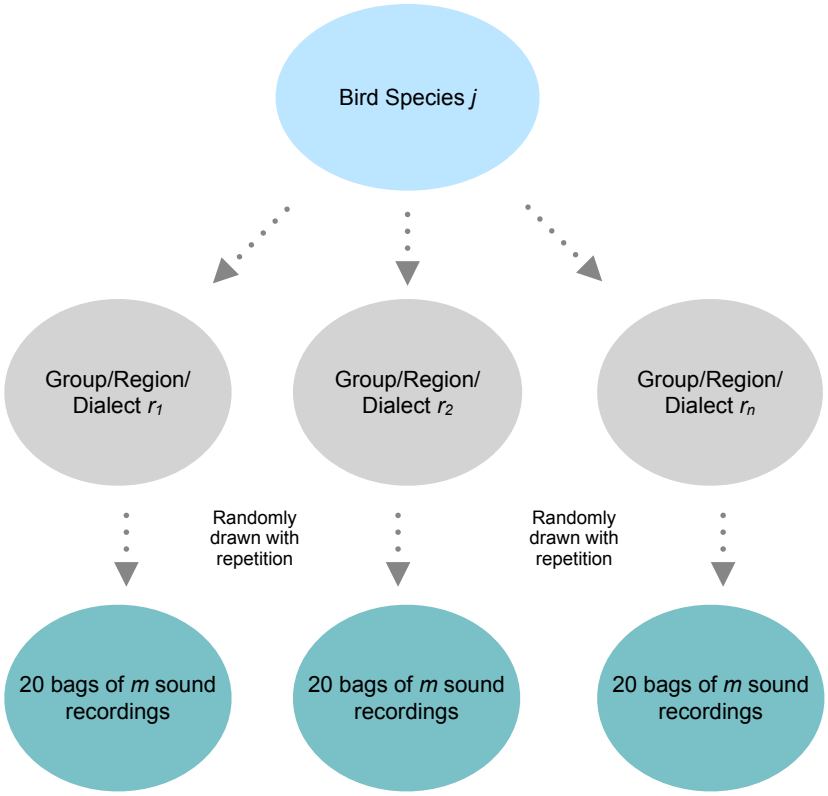


Figure 6.6: Schematic for the *bags-of-birds* approach (Ghani and Hallerberg, 2021). Recording samples for different regions and groups in the case of House Wren and different dialects or dialect groups in the case of Yellowhammer are first organised into larger datasets. Subsequently, for classification, 20 bags or ensembles of recordings are drawn randomly from these larger datasets with repetition to repeat each computational run 20 times.

tuning such that the chances of over-fitting are reduced. In the random forest model, *max_depth* was set to 5 to reduce the complexity of the model. *min_samples_leaf* was set to 2. *n_estimaters* was set to 200. For the rest of the parameters, default values set by the *sklearn* function for the random forest were chosen, since they are shown to provide fairly good estimates. In the case of ANN, the only parameter that we tuned was the learning rate using the keras tuner. The rest of the model parameters were hand-tuned by trial-and-error and the ANN was constructed using one hidden layer, an input layer, and an output layer. The number of neurons in the hidden layer was set to 16.

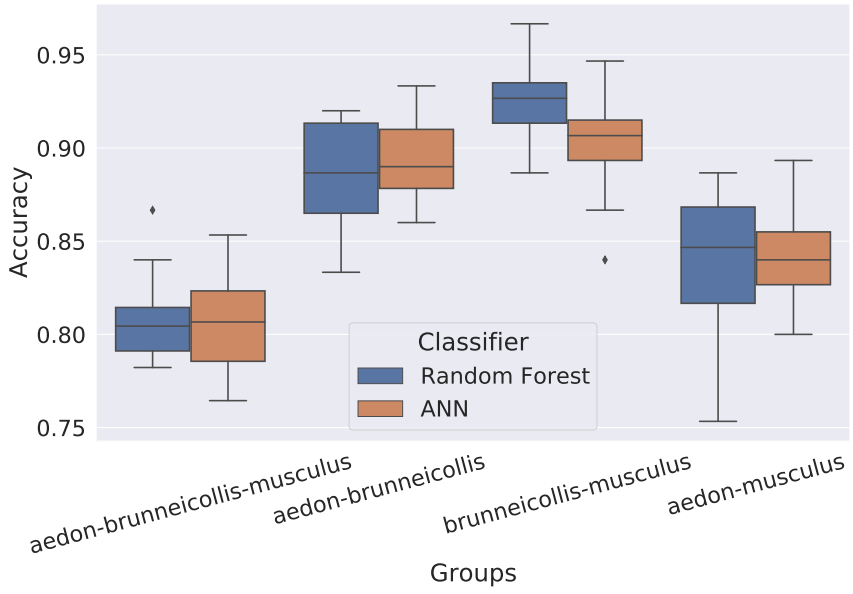
6.3 Results and Discussion

6.3.1 House Wren

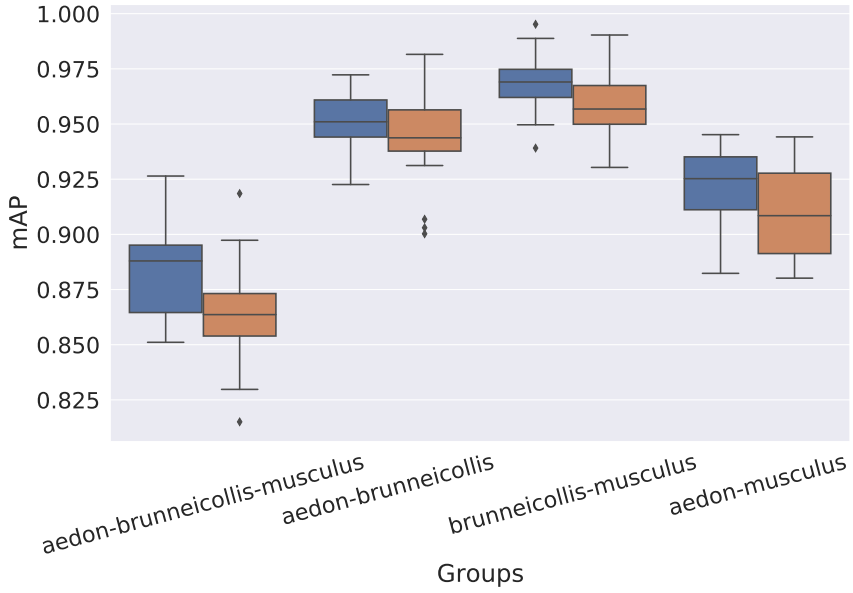
The performance of classification is evaluated by testing for 20 trials drawing 300 recordings from different groups, with repetition, for each trial from the larger datasets. Since datasets for different regions/groups vary in the number of recordings available, this allows us to create balanced datasets while varying the sets of samples in each trial to arrive at a reliable estimate of expected performance. The performance is evaluated using two commonly used metrics, namely, accuracy and mean average precision (mAP). In the case where the classification is tested for three groups of House Wren, namely, the *aedon* group, the *brunneicollis* group, and the *musculus* group, we performed a multi-class classification where the classifier is trained on three classes, and in the second case, we performed a binary classification to train for different combination of the three groups. The results of different classification configurations are summarized using box plots in Fig. 6.7.

There are several aspects of the results that are worth discussing. It can be observed from both the accuracy and mean average precision plots that classifying three groups is, on an average, outperformed by the binary classification between any combination of two groups, since the medians of the box plots are higher for the two groups classification than the median of performance indices for the three groups. This is an expected result since increasing the number of classes leads to a dip in performance as investigated in detail in the previous chapter. The confusion matrices (see Fig. 6.8), averaged over the 20 trials, were plotted to investigate the patterns of misclassifications across groups. For instance, it can be observed in the confusion matrix for three groups that, among the three groups, the *brunneicollis* group is the easiest to distinguish. Firstly, this group has the highest number of true positives and secondly, most of the wrongly classified samples for the other two groups are not classified as belonging to this group. For instance, out of the 75 samples for the *musculus* group, only 2 are labeled by the classifier as *brunneicollis* while 8 are labeled as *aedon*. The binary classifications were carried out to further investigate the influence of individual groups on classification performance. Consequently, the same pattern was observed – out of the three combinations of binary classifications, the ones involving emphbrunneicollis group performed better. The median mAP score is 0.925 using a random forest classifier for the classification involving the *aedon* group and the *musculus* group while the median mAP scores for the other two combinations are about 0.95 and 0.965, respectively.

The classification results were also mapped to the locations where the recordings were made to explore the possibility of some kind of patterns in misclasssifications.



(a)



(b)

Figure 6.7: Classification results using the performance measures a) Accuracy and b) mAP for different combinations of groups for the species House Wren using two machine learning models – a random forest and an artificial neural network. Shown are the ranges from best result to worst result (whiskers) obtained for 20 different randomly drawn subsets of samples. The black marking in each box represents the median and the boxes indicate the middle 50% of the results.

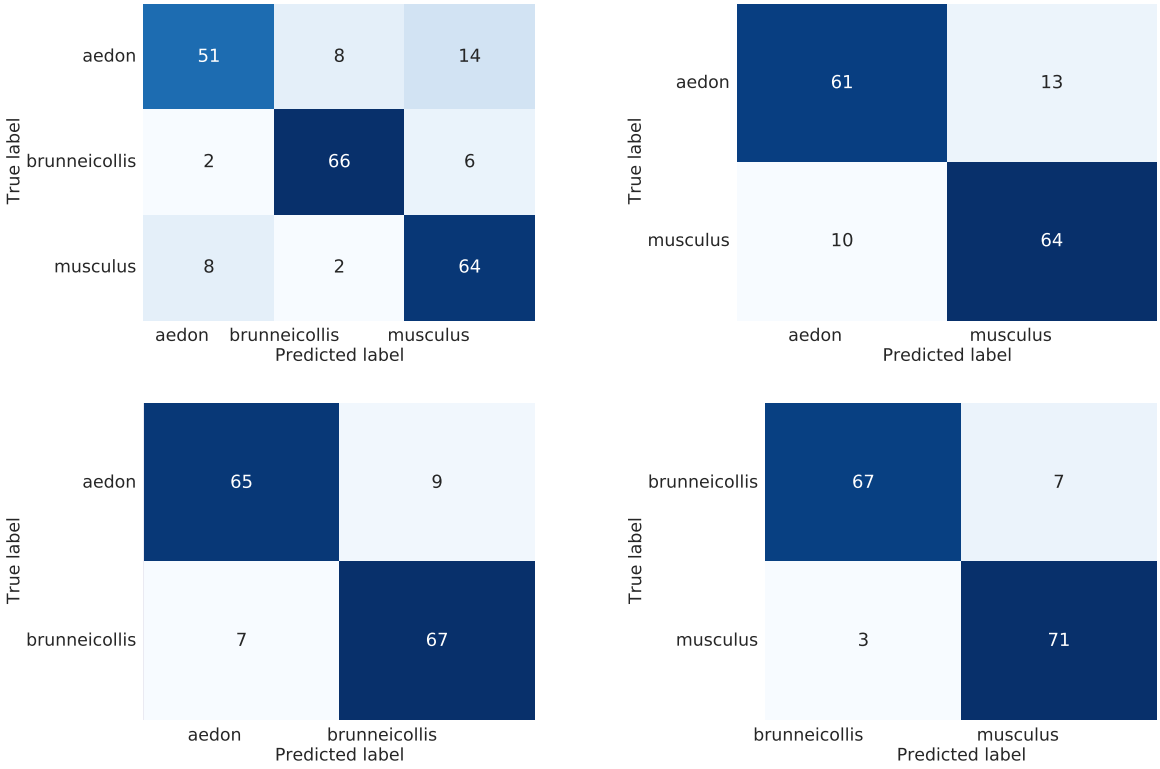


FIGURE 6.8: Confusion Matrices of classification results using a random forest classifier for different combinations of groups for the species House Wren. Note that the confusion matrices plotted here display the average scores over 20 trials to understand the relevant patterns. The upper left figure corresponds to a multi-class classification with three groups – *aedon* group, the *brunneicollis* group, and the *musculus* group. The upper right figure corresponds to a binary classification between the *aedon* group and the *musculus* group. The lower left figure corresponds to a binary classification between the *aedon* group and the *brunneicollis* group and the lower right figure corresponds to a binary classification between the *brunneicollis* group and the *musculus* group. The total number of testing samples used for each group is 75.

It would be interesting if, for instance, the wrongly classified samples were restricted to specific areas. Fig. 6.9 shows plots for the three groups where the blue markers are true positives i.e., the correctly classified samples, and the red markers are false positives i.e., samples belonging to other two groups that have been wrongly classified as this group. Apart from the fact that most misclassifications are accumulated in groups other than the *brunneicollis* group, there is not much of an area-specific pattern that can be observed in the plots.

As per the accuracy metric, the two classifiers have performed almost similarly. Apart from the *brunneicollis-musculus* classification, where the median over 20 trials

is higher using the random forest classifier, the two classifiers provide comparable performance in the other configurations. Using the average mean precision as a measure of classification performance, one can assert that the random forest provides a better score, on an average over 20 trials, compared to the shallow feed forward ANN. The median mAP scores over 20 trials are higher in all the four classification configurations.

Fig. 6.10 provides a summary of the results obtained for the experimental setup where the classifier is trained on the samples of recordings made in 5 different latitudinal zones spanning about 75° of latitude across the Americas. It is evident from the figure that the random forest trained on MFCC coefficients performs better than the ANN in terms of performance of classification. The median mean average precision score for 20 trials for a random forest classifier is about 0.73 while the score using an ANN is 0.59. Therefore, the random forest outperforms the ANN by more than 10% in terms of classififying samples from different latitudinal areas.

The performance drops compared to the group-based classification discussed above. Apart from the fact that the number of classes have increased by 2, by employing latitude-based demarcation, we have essentially divided the *musculus* group into three regions since this is also the largest among the three groups (see Fig. 6.5), and is composed of 20 House Wren sub-species, compared to the other two groups that include between 1-3 sub-species each. This also, therefore, adds to the complexity of classification. Nevertheless, the high number of sub-species perhaps also explains the possible variablility in vocalizations within this group and, consequently, the relatively high classification performance for these regions. We can see from the averaged confusion matrix in Fig. 6.11 that more than half of the testing samples are correctly classified in the lower three regions and, in the region spanning between -15°-0° latitude, about two-thirds of the samples have been correctly classified.

Given that there have been relatively few studies focusing on latitudinal variation in songs variability on an intraspecific level (Kaluthota et al., 2016), these results can be useful to biologists that are trying to understand the evolutionary divergence in vocalizations or study the existence of sub-species and dialects. For instance, such an audio-based analysis can provide hints to biologists on where to look for interesting patterns. The results also show that it is possible to extract such information on vocal variations based on a solely data-driven approach using recordings of vocalizations.

A similar pattern for inter-region misclassifications can bee seen here as was observed in the inter-group classification (see Fig. 6.8). Most misclassifications of the region spanning 30°-45°, which corresponds to the *Troglodytes aedon* group of House Wren sub-species, are accumulated in the lower most three regions (corresponding to the *brunneicollis* group). Now that the misclassifications can be seen in more resolution, most of the wrongly classified samples can be seen as belonging to the region

FIGURE 6.9: Classification results for three groups of House Wren mapped to locations where the respective sound samples were recorded. The upper left map corresponds to the Northern House Wren (*Troglodytes aedon* group). On the upper right side is the map for *Troglodytes brunneicollis* group. The lower map shows result for the Southern House Wren (*Troglodytes musculus* group). The blue markers correspond to correct classifications (true positives) and red markers correspond to incorrect classifications (false positives).

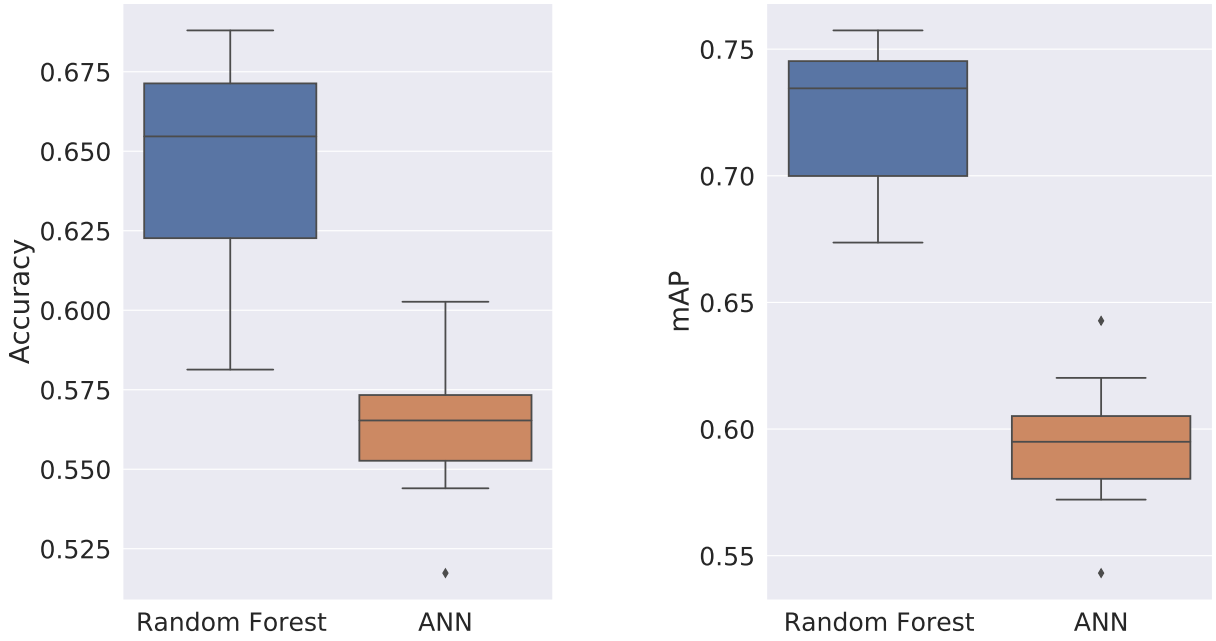


Figure 6.10: Classification results using the performance measures a) Accuracy and b) mAP for samples from 5 regions obtained by controllling for latitude between 45°N and 30°S for the species House Wren using two machine learning models – a random forest and a artificial neural network trained on MFCC coefficients. Shown are the ranges from best result to worst result (whiskers) obtained for 20 different randomly drawn subsets of samples. The black marking in each box represents the median and the boxes indicate the middle 50% of the results.

spanning -15°-0° latitude. Moreover, a significant proportion of samples for the lower most two regions are misclassified with the label for the region spanning the 30°-45° latitude region. It also seems, from the confusion matrix, that the samples in regions spanning 30°-45° and -15°-0° latitude are more similar to each other compared to the samples in 0°-15° regions. Once again, the region spanning 30°- 45° latitude (also the *brunneicollis* group) has the highest number of correctly classified samples, similar to the result of the previous experiment.

This pattern of increased misclassifications between the two extreme latitudinal ranges or between *aedon* and *musculus* groups is perhaps consistent with the conventional prediction that metrics of song complexity at the higher latitudes in the Southern Hemisphere might converge on the song metrics in the northern temperate zone. The hypothesis avers that populations in higher latitudes in Southern Hemisphere face functional pressures on songs due to strongly seasonal environments similar to populations in the Northern Hemisphere (Kaluthota et al., 2016). The similarities

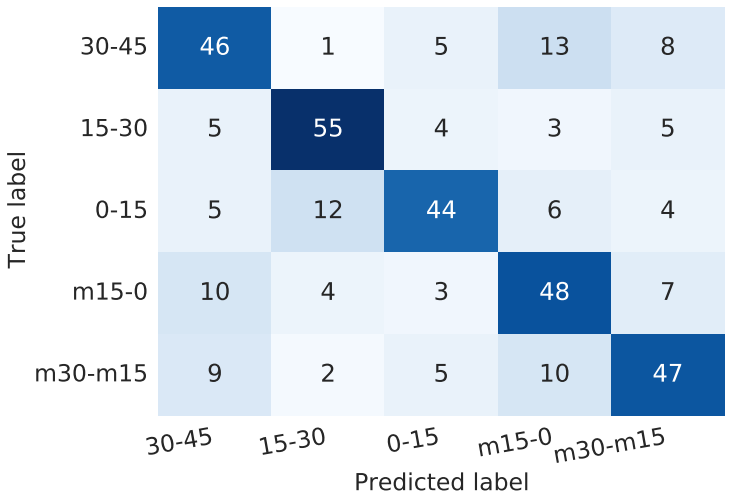


FIGURE 6.11: Confusion Matrix of classification results using a random forest classifier for samples from 5 regions obtained by controlling for latitude between 45°N and 30°S for the species House Wren. Note that the confusion matrix plotted here displays the average scores over 20 trials to understand the relevant patterns. The total number of testing samples used for each region is 75. m in the latitude-range labels stands for minus.

in seasonal environments can be, for instance, in the average temperatures, the day length etc. It was also documented in Kaluthota et al. (2016) that the male House Wrens in higher latitudes in both the Southern and Northern Hemispheres produced more elements per song by singing at higher rates rather than extending their song lengths. They found a U-shaped quadratic relationship for 14 song metrics of latitudes, varying it between high latitudes in the two hemispheres.

It is worthwhile to mention that, compared to interspecies classification, intraspecies classification is a more challenging task for a classifier since the songs within species can sometimes show variations in a very minute fraction of the song. For instance, differences can be on a syllable level while most of the song pattern stays the same. This must have an influence on the predictive power of the features employed in this work. To elaborate further, since the span of differing samples within a recording is very small, it is likely that due to the nature of segments joining process employed here to construct 5-second recordings, there are recordings that do not include the differing parts of songs. These results can be improved in future works by overcoming such scenerios. Furthermore, a lot of misclassifications are a result of noisy samples or overlapping sounds of same or even different birds species and other living organisms.

6.3.2 Yellowhammer

The performance of classification is evaluated by testing for 20 trials, drawing samples for each group with repetition, for each trial from the larger datasets. Since datasets for different dialects vary in the number of recordings available, this allows us to create balanced datasets while varying the sets of samples in each trial to arrive at a reliable estimate of expected performance. In this way, we can also provide estimates of spread around median scores. The performance is evaluated using two commonly used metrics, namely, accuracy and mean average precision (mAP). Two different machine learning models, similar to the ones for the analysis of House the Wren, are employed to carry out classification.

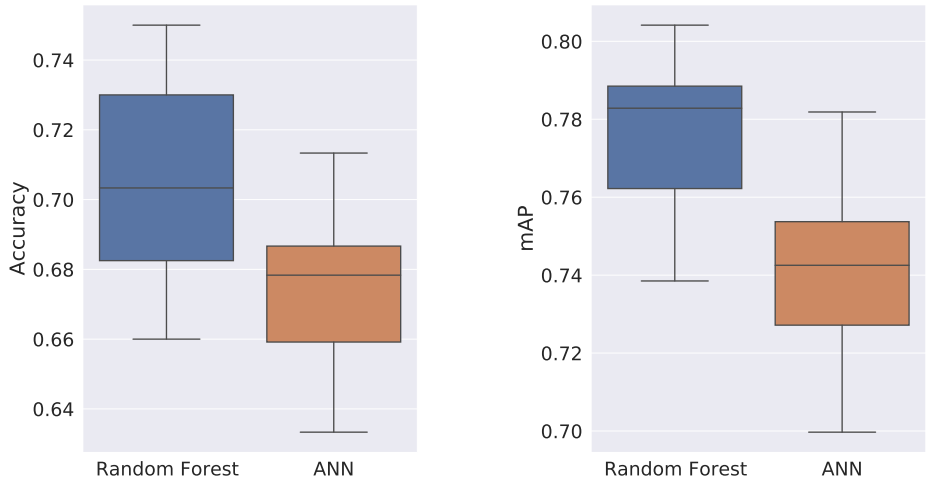


Figure 6.12: Classification results using the performance measures a) Accuracy and b) mAP for two dialects groups of Yellowhammer sounds, the *B* group, and the *X* group, using two machine learning models – a random forest and a artificial neural network trained on MFCC coefficients. Shown are the ranges from best result to worst result (whiskers) obtained for 20 different randomly drawn subsets of samples. The black marking in each box represents the median and the boxes indicate the middle 50% of the results.

The classification is performed on two levels of complexity for different dialects of the Yellowhammer species. In the first experiment, the classification is performed for two groups of dialects, namely, the *B* group and the *X* group.

Results of the classification for the two groups of Yellowhammer sounds are summarized using box plots in Fig. 6.12. We observe, at the outset, that random forest performs better compared to the shallow feed forward ANN. The median scores on both metrics are higher using the random forest. The spread of scores for different subsets of recordings is rather large. Difference in mean average precision, for

instance, between the lowest and highest scores is about 9% using a random forest classifier. The ANN results are no different. This shows that the choice of recordings has a considerable influence on the performance of classification.

The confusion matrix (see Fig. 6.13), averaged over the 20 trials, was plotted to investigate the pattern of misclassifications for the two groups of dialects. The confusion matrix is generated for the random forest classifier. For instance, it can be observed in the confusion matrix that the classifications for the *X* have been outperformed by the *B* group in terms of number of misclassifications in each group. Given the complexity of the problem, the classification results are promising since the dialects only vary in the last two syllables of the entire song. The first part of the song is shared by all dialects as can be seen in Fig. 6.4.

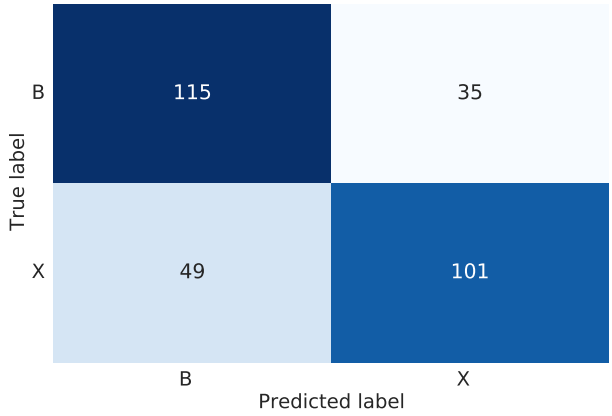
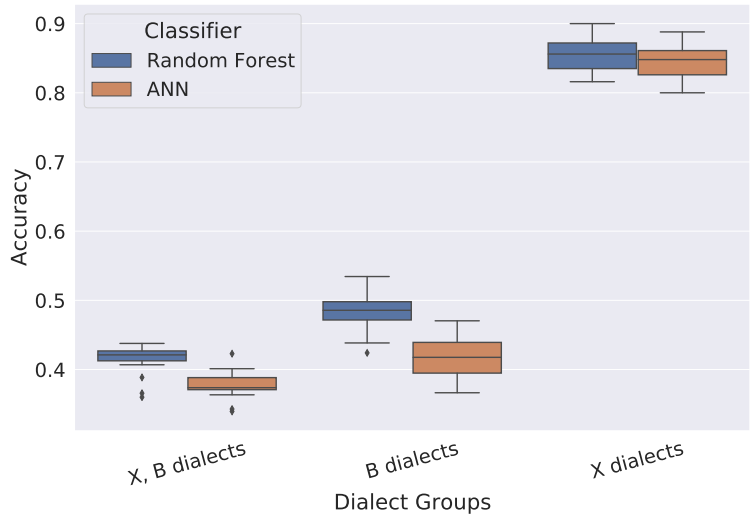


Figure 6.13: Confusion Matrix of classification results using a random forest classifier for two dialects groups of Yellowhammer sounds, the *B* group and the *X* group. Note that the confusion matrix plotted here displays the average scores over 20 trials to understand the relevant patterns. The total number of testing samples used for each region is 150.

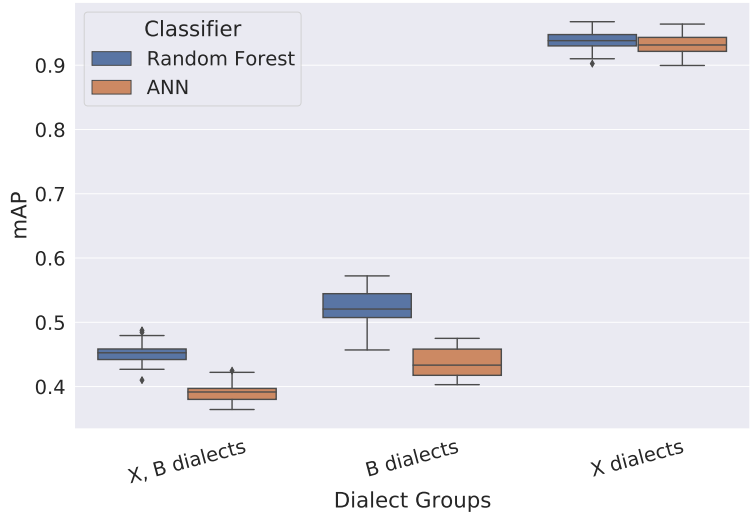
In the second configuration, the classification is performed on the higher level of complexity in which all 7 different dialects are considered. Apart from the classification involving all dialects, we also perform two separate classifications consisting of different variations of the *B* group and different dialects in the *X* group. The results of the classification for the different configurations are summarized using box plots in Fig. 6.14.

It can be observed in the figure that including all dialects significantly lowers the performance of classification. The random forest, once again, performs a little better than the ANN. These results are not surprising for several reasons.

Firstly, the Yellowhammer recordings contain a lot of unfinished songs which makes the whole task of distinction highly challenging since the last part is the only



(a)



(b)

Figure 6.14: Classification results using the performance measures a) Accuracy and b) mAP for 3 different configurations of Yellowhammer dialects grouped together, using two machine learning models – a random forest and an artificial neural network trained on MFCC coefficients. *X, B dialects* incluces all 7 Yellowhammer dialects found in Czech Republic. *B dialects* includes the 5 dialects where the second last syllable is *B*. *X dialects* includes the 2 dialects where the second last syllable is *X*. Shown are the ranges from best result to worst result (whiskers) obtained for 20 different randomly drawn subsets of samples. The black marking in each box represents the median and the boxes indicate the middle 50% of the results.

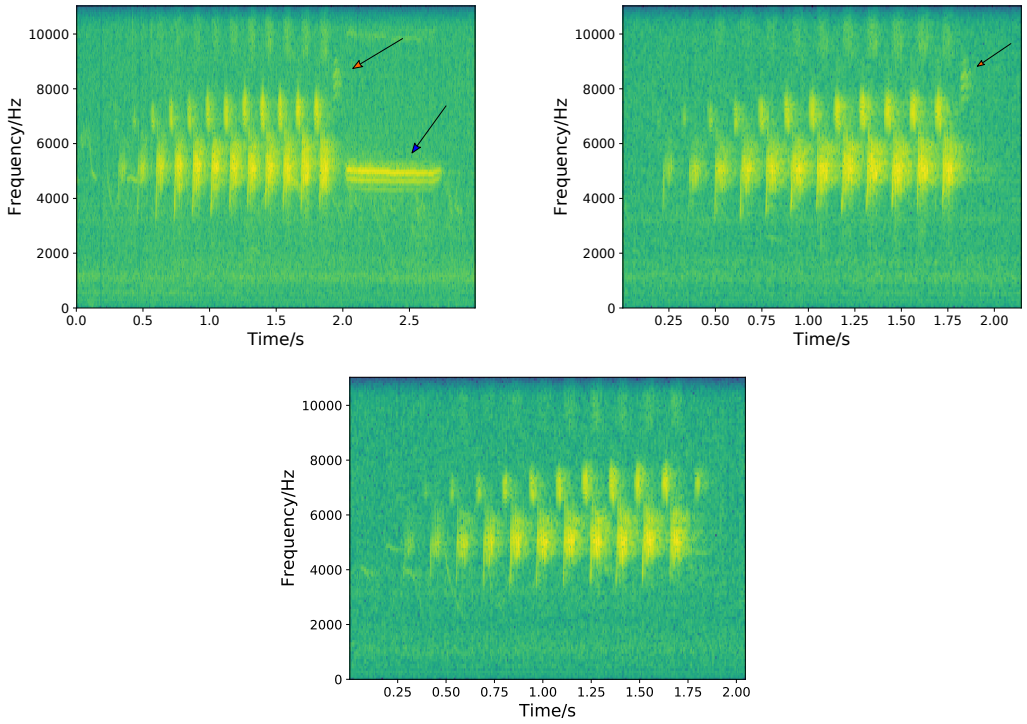


FIGURE 6.15: Spectrograms of three instances of a dialect *XsB* of a Yellowhammer song extracted from a single recording. The upper left figure corresponds to a finished song. On the upper right side is the spectrogram for a partly finished song. The lower figure is the spectrogram for a fully unfinished song. The red and blue pointed arrows correspond to the second last and the last syllables, respectively. The x and y axes correspond to time (in seconds) and frequency (in Hertz).

distinguishing characteristic for separating different dialects. There are partly unfinished songs and fully unfinished ones. The former involves those songs that include the first of the last two syllables and the fully unfinished songs do not include either of the last two syllables. Fig. 6.15 shows spectrograms for three segments extracted from a single recording of a Yellowhammer song with *XsB* dialect. The upper left spectrogram shows the full song, the upper left shows a partly unfinished song with the last syllable missing, and the lower one shows a fully unfinished song with the last two syllables missing.

Secondly, as discussed before, the dialects vary for a very short proportion of the song just towards the end. Each song is about 2.5s long and the last two syllables, the different combinations of which give rise to different dialects, range between 0.5-1.0s in length. Besides, the individual syllables within *X* group and *B* group only vary in one syllable.

Thirdly, the differences are in some cases so minute that deciding a threshold to

separate two dialects is a challenging task. Such is the case, for instance, for dialects *BD* and *BlBh*. Researchers working on Yellowhammer dialects at the Charles University of Prague found it challenging to find a boundary that could separate these two dialects and had to set some arbitrary threshold from the distribution of frequencies for different samples. Similarly, to separate the dialects *BE* and *BD* was challenging, because the *D* syllable in *BD* dialect is not always flat and has a rather ascending frequency slope in some cases, similar to the syllable *E* in *BE*. Here again, these researchers had to set arbitrary thresholds on the slope to separate the samples.

The averaged confusion matrix for the classification for all dialects, in Fig. 6.16, shows that, among all dialects, *XsB* and *BlBh* are the easiest to classify while *BhBl* is the most difficult.

We observe a slightly better performance for the classification for 5 different *B* dialects compared to classifying all dialects. Here again, we notice from the confusion matrix that *BlBh* is the easiest to classify. Given that there are significant numbers of both partly unfinished and fully unfinished songs in the recordings, this is not a surprising result. Since most of the B dialects only vary in the last syllable of the terminal phrase, which means in either of the two scenerios – partly unfinished songs and fully unfinshed songs – the distinguishing parts for such dialects are simply absent. Therefore, the classification performance for this dialect group takes a hit.

The *X* dialects, on the other hand, have performed surprisingly well with a median mAP score of about 0.94. The fact that there are only two classes to separate is also a contributing factor that makes this case less challenging. Nevertheless, such a high performance perhaps also implies that separating these two dialects is relatively easy. The same hypothesis provided in the previous case can perhaps also explain this result. Since X group varies in the first syllable of the last phrase, this syllable can also be encapsulated by the party unfinished songs as can be observed in Fig. 6.15. Therefore, there are more recordings that can capture the difference between the two dialects in the X group. It can also be observed in the confusion matrix for all dialects that misclassifications among these two dialects are also minimal. Furthermore, it seems that between the two dialects, the *XlB* dialect is easier to separate, given the classifier, on an average, wrongly classifies only 9 times compared to 25 times for the other dialect.

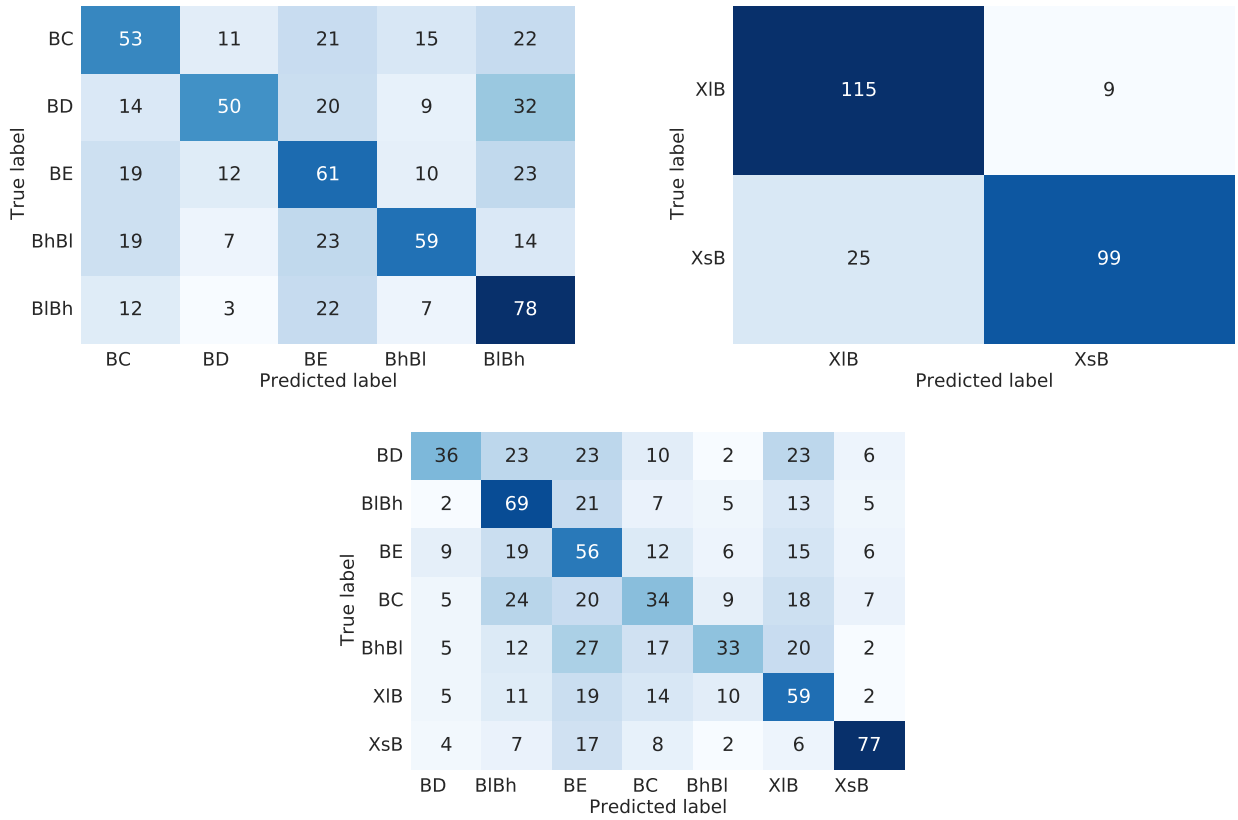


FIGURE 6.16: Confusion Matrices for classification results obtained using a random forest classifier for 3 different configurations of Yellowhammer dialects grouped together, the *B dialects* (upper left), the *X dialects* (upper right), and the *X, B dialects* (lower). Note that the confusion matrix plotted here displays the average scores over 20 trials to understand the relevant patterns. The total number of testing samples used for each region is 125.

Conclusions 7

"It's not only the question, but the way you try to solve it."

Maryam Mirzakhani

An automated analysis of bird vocalizations allows for a cost and effort-effective way to gain insights about different aspects concerning birds in heavy contrast to conventional observer-based surveying methods. One aspect that is of foremost interest today is monitoring of bird species to gather reliable estimates of population trends, sizes, and ranges for different species around the world. While the information and knowledge gained from such an assessment would assist in efforts that are directed toward conservation of species that are at risk of extinction, it would also provide us cues about the health of the ecosystem in general. Passive acoustic monitoring has made it possible to overcome spatial and temporal limitations presented by classical observer-based methods. Nevertheless, this has posed new challenges of sifting through huge amounts of audio data gathered using autonomous recording units. The need for automated methods that can provide reliable estimates of informative trends and yet be cost-effective in terms of computational constraints is of key interest to realize the full potential of passive acoustic monitoring. In this thesis, several aspects of automated classification of bird species based on their vocalizations

are investigated with the goal of furthering the understanding of challenges that need to be addressed. The key findings of this thesis are summarized in the next sections.

7.1 On Relying on Simple, Computationally Inexpensive Frameworks

1. We investigated the predictive power of a simple statistical technique to compare estimated distributions from different bird species and their mean cepstra, and found that advanced machine learning models can provide considerably better performance for data that is complex such as bird sounds. A random forest classifier was trained with pre-computed features and the performance obtained on 10 species was comparable to several state-of-the-art classification techniques. A median mean average precision of about 0.84 was obtained across 20 subsets of 10 bird species that were drawn randomly. The different species of birds were not purposely chosen as has been seen in many previous works. Therefore, we can safely propose that the classification framework employed in this work can be employed for all types of bird species.

2. Several aspects of audio-based analysis of bird species were studied. The predictive power of different commonly employed features was compared. It was found that mel-scaled features performed better than linearly scaled ones. MFCCs, by far, led to the highest classification performance when first and second sample moments were used to summarize the time series of feature coefficients. Additionally, it was observed that segmenting signal parts from the field recordings in order to discard the noisy parts led to an increase in performance. Since, until this point, both vocalization types – calls and songs – were used to classify bird species, we also performed a computational experiment to study the influence of individual sound types on the classification performance. It was observed that training a model on a single vocalization type led to higher classification performance. The results of this experiment showed that it was easiest to classify birds based on their calls. Furthermore, the influence of the quality of sound recordings was investigated. It was found that high quality recordings did lead to a higher performance.

3. We have employed a random forest classifier trained on pre-computed features in order to analyze the classification performance using simple, computationally inexpensive and yet tremendously powerful models that can be trained on relatively less amounts of data as opposed to very deep neural networks.

7.2 On Dependence of Classification Performance on the Number of Bird Species

1. In order to estimate the robustness of the classification approach, we introduce a novel *bags-of-birds* approach. Following this approach, the bird species are not purposefully chosen but 20 bags or ensembles of bird species are constructed where the species are drawn randomly with repetition from the complete dataset of 659 bird species. The idea is to come up with reliable estimates of classification performance by repeating the computations 20 times. In this way, we are also able to provide the mean and spread for the expected performance on the machine learning procedure.

2. We investigated the dependence of classification performance on the number of species. This was done by randomly drawing n species from the larger dataset where n was varied between 10 and 300. For each n, the computations were repeated 20 times on balanced subsets. The classification performance was measured using several metrics. This was done primarily since there has been no consensus in the past works on the choice of evaluation metric employed. This makes it difficult to compare results across different works. We found that all metrics for classification performance showed a decline as the number of species were increased. For some metrics, the trend was explained analytically knowing the n-dependence of confusion matrix and making an assumption of an idealized perfect classifier. We again relied on a shallow feed forward neural network to benchmark the performance using models that demand minimal computational power and can work with smaller datasets. We found that the model provides a comparable performance when the species size is not that large, but for large species' sizes exceeding, say, 100, the deep neural network architectures have shown a considerably better performance. Given that we high classification performance can be obtained using less data hungry and less computationally complex methods for datasets with moderate amount of species, it can be inferred that employing very deep networks for such tasks is simply an overkill.

3. We observed that classification performance, apart from the number of species considered, also depends on the individual composition of bird datasets. The interquartile-ranges for the classification results varied as much as 12% for the same value of n between two different *bags-of-birds*. For this reason, we assert that the generic claims about performance for, say, n number of species should not be interpreted as a measure of performance for any n species. Since the classification performance is heavily contingent on the species selected, it might

not generalize to another set.

4. We analyzed the effect of introducing a confidence threshold on the minimum probability required by a class for a sample to be assigned to that class. This was done to measure if the level of confidence, low or high, with which the classifier assigns labels to samples matches the frequencies with which the classification model classifies these samples. We observe that the precision increases as the confidence threshold is increased which shows that when the model is assigning high confidence to its classifications, the classifications are mostly correct which would be expected of a good classifier. Nonetheless, for the samples the model is not that confident about (lower confidence thresholds), it errs on the side of over-confidence.

7.3 On Quantifying Intra-Species Variability in Songs in Widespread Species

1. We extend the analysis from inter-species to intra-species classification. While several results have been reported for classification and recognition of different bird species, using machine learning models to classify variations within the bird species has not been explored. Biologists have been relying on manual approaches to investigate such phenomena. We have considered two bird species, namely, the House Wren and Yellowhammer. Both of these species are widely spread across vast geographical areas. In case of House Wren, we carried classification to explore possible variations in songs for different groups of subspecies and different latitudinal regions and for Yellowhammer, the task was to classify 7 established dialects found in the Czech Republic.

2. For House Wren, classification is first performed for three groups of subspecies, namely, the *aedon*, *brunneicollis*, and *musculus* groups that have been reported to have developed independent evolutionary trajectories. Two models – a random forest and a feed forward neural network – are trained on MFCC features. Both the models classify the three groups with a high performance. To further investigate the inter-group vocal variations, binary classifications were performed between combinations of three groups. It was observed from all these computational experiments, that the *brunneicollis* group was the easiest to distinguish while most misclassifications happened between the *aedon* and *musculus* groups. We could find some evidence for this trend in the literature which suggests that populations in the higher latitudes in the two hemispheres face similar functional pressures on songs due to strongly seasonal environments. A latitudinal

region-based classification of House Wren sounds was also conducted, where the area between $45°$N and $30°$S is divided into 5 regions and songs recorded in these regions are classified. We wanted to confirm the studies that have reported latitudinal variations in House Wren species in the Americas. Our model classified the recordings from different regions with a median mean average score of about 0.73 across 20 runs for different sets of recordings drawn with repetition.

3. For Yellowhammer, we performed the classification broadly on two levels of complexity for different dialects. In the first case, we classified two groups: B dialects and X dialects. In the second case, we set up three configurations: classification of all 7 dialect types, classification of B dialect types, and classification of X dialect types. We observed that both models classified the B and X groups with relatively high performance. The random forest provided a higher score. We obtained a median mean average score of about 0.79 across 20 runs for different sets of recordings. For the different configurations in the second case, classifying all 7 dialects led to the lowest performance, which was followed by classying different B dialect types. Classifying the two X dialect types led to the highest classification performance across all setups.

4. In this way, we could confirm intraspecific vocal variation in bird species based solely on a data-driven approach. Moreover, following the results obtained, we can infer that automated classification using machine learning is capable of replacing manual approaches to classify vocal divergence and different dialects in a way that does not require hundreds or thousands of hours of manual inspection of recordings to measure spectral characteristics of songs. Such an approach can assist biologists by providing them hints on where to look for interesting trends that could help further our understanding of the evolution of variations in bird vocalizations.

7.4 Outlook

There are some ideas that could not be pursued due to paucity of time but can form a basis for future work towards improving the tasks carried out in this thesis. A list of some promising directions are presented as follows:

1. Unsupervised feature learning can be explored as a means of finding transformations in data that are driven by localized properties in the data. Learning features that will allow extracting the most descriptive variations inherent in target signals can help us overcome the selectivity-invariance problem and provide a trade-off between computationally expensive deep architectures and con-

ventional machine learning algorithms that rely on manually designed features. Feature learning has produced great success in the field of computer vision and its effectiveness can be explored for the task of automated analysis of bird sound recordings.

2. We observed that noise (low SNR data) has a significant influence on classification performance. Developing better techniques aimed at denoising field recordings before feeding these recordings into the classification system can improve predictions.

3. In addition to choosing features and models that exploit the discriminatory variations in sound data, temporal context present in bird vocalizations can be used to more accurately classify different sounds. This can improve classification especially in case of dialects or intraspecific variation in songs where differences can be very minute and meagre. Madhusudhana et al. (2021) have recently reported improved automatic recognition on inclusion of temporal information in the case of fin whale songs.

4. The data on Yellowhammer dialects from other countries, apart from the Czech Republic, can be used to reproduce the results obtained in this thesis. Similarly, for House Wren, the variations in songs across different regions and groups can be tested using an additional dataset provided by a different source to verify the findings.

5. In the case of Yellowhammer dialects, the classifier can be trained to classify between finished songs, partially finished songs, and fully finished songs.

6. In this thesis, the datasets employed have included focal recordings with single species or at least the annotated species forming the most prominent audioprint in the respective recordings. In the future, models can be trained to perform on soundscapes that can include multiple species vocalizing in a single recording or even overlapping sounds from different species. This is one of the biggest challenges in automated classification of bird sounds and the performance provided by the deepest and most complex neural architectures remains scant to date.

7. Instead of traditional backpropagation the artificial neural network can be trained using Bayesian inference (Jospin et al., 2020). Such models explicitly incorporate uncertainty within its formulation by assigning probability distributions to weights and employing probabilistic activation functions with the intent of getting better insights on the uncertainties associated with underlying processes.

Bibliography

The iucn red list of threatened species, 2021. URL `https://www.iucnredlist.org/resources/summary-statistics`.

Xeno canto, https://www.xeno-canto.org/, accessed <2020>.

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.

Gebremedhin Teklemariam Abreha. An environmental audio-based context recognition system using smartphones. Master's thesis, University of Twente, 2014.

AOU. Aou checklist of north and middle american birds, 1998.

Gregory F Ball and Stewart H Hulse. Birdsong. *American Psychologist*, 53(1):37, 1998.

Ednei Barros dos Santos et al. *The nature and function of song diversity in southern house wrens (Troglodytes aedon chilensis)*. PhD thesis, Lethbridge, Alta.: Universtiy of Lethbridge, Department of Psychology, 2018.

Selin Bastas, Mohammad Wadood Majid, Golrokh Mirzaei, Jeremy Ross, Mohsin M Jamali, Peter V Gorsevski, Joseph Frizado, and Verner P Bingman. A novel feature extraction algorithm for classification of bird flight calls. In *2012 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1676–1679. IEEE, 2012.

Selin A Bastas. *Nocturnal bird call recognition system for wind farm applications*. PhD thesis, University of Toledo, 2012.

George Bebis and Michael Georgiopoulos. Feed-forward neural networks. *IEEE Potentials*, 13(4):27–31, 1994.

Michael J Bianco, Peter Gerstoft, James Traer, Emma Ozanich, Marie A Roch, Sharon Gannot, and Charles-Alban Deledalle. Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, 146(5): 3590–3628, 2019.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

David Brewer. *Wrens, dippers and thrashers*. Bloomsbury Publishing, 2010.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

Clive K Catchpole and Peter JB Slater. *Bird song: biological themes and variations*. Cambridge university press, 2003.

Juan Gabriel Colonna, Marco Cristo, Mario Salvatierra Júnior, and Eduardo Freire Nakamura. An incremental technique for real-time bioacoustic signal segmentation. *Expert Systems with Applications*, 42(21):7367–7374, 2015.

Jason Cramer, Vincent Lostanlen, Andrew Farnsworth, Justin Salamon, and Juan Pablo Bello. Chirping up the right tree: Incorporating biological taxonomies into deep bioacoustic classifiers. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 901–905. IEEE, 2020.

RB Cunningham, DB Lindenmayer, and Bruce D Lindenmayer. Sound recording of bird vocalisations in forests. i. relationships between bird vocalisations and point interval counts of bird numbers–a case study in statistical modeling. *Wildlife Research*, 31(2):195–207, 2004.

Theodoros Damoulas, Samuel Henry, Andrew Farnsworth, Michael Lanzone, and Carla Gomes. Bayesian classification of flight calls with a novel dynamic time warping kernel. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 424–429. IEEE, 2010.

Charles Darwin. *The descent of man, and selection in relation to sex*, volume 2. D. Appleton, 1872.

Josep del Hoyo. *Handbook of the birds of the world. 6. Mousebirds to Hornbills*. Lynx Ed., 2001.

Lucie Diblíková, Pavel Pipek, Adam Petrusek, Jiří Svoboda, Jana Bílková, Zdeněk Vermouzek, Petr Procházka, and Tereza Petrusková. Detailed large-scale mapping of geographical variation of yellowhammer emberiza citrinella song dialects in a citizen science project. *Ibis*, 161(2):401–414, 2019.

Seppo Fagerlund. Automatic recognition of bird species by their sounds. *Finlandia: Helsinki University Of Technology*, 2004.

Seppo Fagerlund. Bird species recognition using support vector machines. *EURASIP Journal on Advances in Signal Processing*, 2007:1–8, 2007.

Almo Farina. *Soundscape ecology: principles, patterns, methods and applications*. Springer, 2013.

Terrence L Fine. *Feedforward neural network methodology*. Springer Science & Business Media, 2006.

D Fink, T Auer, A Johnston, M Strimas-Mackey, O Robinson, S Ligocki, B Petersen, C Wood, I Davies, B Sullivan, et al. ebird status and trends, data version: 2018; released: 2020. *Cornell Lab of Ornithology, Ithaca, New York*, 2020.

Elizabeth JS Fox, J Dale Roberts, and Mohammed Bennamoun. Text-independent speaker identification in birds. In *Ninth International Conference on Spoken Language Processing*, 2006.

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Brett J Furnas. Rapid and varied responses of songbirds to climate change in california coniferous forests. *Biological Conservation*, 241:108347, 2020.

Burooj Ghani and Sarah Hallerberg. A randomized bag-of-birds approach to study robustness of automated audio based bird species classification. *Applied Sciences*, 11(19):9226, 2021.

Theodoros Giannakopoulos and Aggelos Pikrakis. *Introduction to Audio Analysis: a MATLAB® approach*. Academic Press, 2014.

I Goodfellow, Y Bengio, and A Courville. Softmax units for multinoulli output distributions. deep learning, 2018.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview, 2020.

Robert M Gray and Lee D Davisson. *An introduction to statistical signal processing*. Cambridge University Press, 2004.

Thibaud Gruber and Laurent Nagle. Territorial reactions of male yellowhammers (emberiza citrinella) toward a specific song structure. *Journal of ornithology*, 151(3): 645–654, 2010.

David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186, 2001.

P Hansen. Geographic song variation in the yellowhammer (emberiza citrinella). *Natura Jutlandica*, 21:209–219, 1985.

Steve NG Howell and Sophie Webb. *A guide to the birds of Mexico and northern Central America*. Oxford University Press, 1995.

Yang Hu and Gonçalo C Cardoso. Are bird species that vocalize at higher frequencies preadapted to inhabit noisy urban areas? *Behavioral Ecology*, 20(6):1268–1273, 2009.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Peter Jančovič and Münevver Köküer. Automatic detection and recognition of tonal bird sounds in noisy environments. *EURASIP Journal on Advances in Signal Processing*, 2011(1):982936, 2011.

Peter Jančovič and Münevver Köküer. Acoustic recognition of multiple bird species based on penalized maximum likelihood. *IEEE Signal Processing Letters*, 22(10): 1585–1589, 2015.

Hohyub Jeon, Yongchul Jung, Seongjoo Lee, and Yunho Jung. Area-efficient short-time fourier transform processor for time–frequency analysis of non-stationary signals. *Applied Sciences*, 10(20):7208, 2020.

Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. Hands-on bayesian neural networks–a tutorial for deep learning users. *arXiv preprint arXiv:2007.06823*, 2020.

S Kahl, T Denton, H Klinck, H Glotin, H Goëau, WP Vellinga, R Planqué, and A Joly. Overview of birdclef 2021: Bird call identification in soundscape recordings. *Working Notes of CLEF*, 2021a.

Stefan Kahl, Thomas Wilhelm-Stein, Holger Klinck, Danny Kowerko, and Maximilian Eibl. Recognizing birds from sound-the 2018 birdclef baseline system. *arXiv preprint arXiv:1804.07177*, 2018.

Stefan Kahl, Fabian-Robert Stöter, Hervé Goëau, Hervé Glotin, Robert Planque, Willem-Pier Vellinga, and Alexis Joly. Overview of birdclef 2019: large-scale bird recognition in soundscapes. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*, number 2380, pages 1–9. CEUR, 2019.

Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61: 101236, 2021b.

Chinthaka Kaluthota, Benjamin E Brinkman, Ednei B Dos Santos, and Drew Rendall. Transcontinental latitudinal variation in song performance and complexity in house wrens (troglodytes aedon). *Proceedings of the Royal Society B: Biological Sciences*, 283(1824):20152765, 2016.

Samina Khalid, Tehmina Khalil, and Shamila Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *2014 science and information conference*, pages 372–378. IEEE, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Anssi Klapuri and Manuel Davy. Signal processing methods for music transcription. 2007.

DE Kroodsma. The diversity and plasticity of birdsong. *Nature's music: the science of birdsong*, pages 108–131, 2004.

DE Kroodsma, D Brewer, J DEL HOYO, A ELLIOTT, and DA CHRISTIE. Handbook of the birds of the world, 2005.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Mario Lasseck. Bird song classification in field recordings: winning solution for nips4b 2013 competition. In *Proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod. org/nips4b, joint to NIPS, Nevada*, pages 176–181, 2013.

Jack LeBien, Ming Zhong, Marconi Campos-Cerqueira, Julian P Velev, Rahul Dodhia, Juan Lavista Ferres, and T Mitchell Aide. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecological Informatics*, 59:101113, 2020.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

Francis F Li and Trevor J Cox. *Digital Signal Processing in Audio and Acoustical Engineering*. CRC Press, 2019.

Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *Ismir*, volume 270, pages 1–11, 2000.

Marcelo T Lopes, Lucas L Gioppo, Thiago T Higushi, Celso AA Kaestner, Carlos N Silla Jr, and Alessandro L Koerich. Automatic bird species identification for large number of species. In *2011 IEEE International Symposium on Multimedia*, pages 117–122. IEEE, 2011.

Irby J Lovette and John W Fitzpatrick. *Handbook of bird biology*. John Wiley & Sons, 2016.

Elizabeth A MacDougall-Shackleton and Scott A MacDougall-Shackleton. Cultural and genetic evolution in mountain white-crowned sparrows: song dialects are associated with population structure. *Evolution*, 55(12):2568–2575, 2001.

Shyam Madhusudhana, Yu Shiu, Holger Klinck, Erica Fleishman, Xiaobai Liu, Eva-Marie Nosal, Tyler Helble, Danielle Cholewiak, Douglas Gillespie, Ana Širović, et al. Improve automatic detection of animal call sequences with temporal context. *Journal of the Royal Society Interface*, 18(180):20210297, 2021.

Peter Marler and Miwako Tamura. Song" dialects" in three populations of white-crowned sparrows. *The Condor*, 64(5):368–377, 1962.

Peter R Marler and Hans Slabbekoorn. *Nature's music: the science of birdsong*. Elsevier, 2004.

Brian McFee, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, D Lee, O Nieto, E Battenberg, D Ellis, R Yamamoto, et al. librosa/librosa: 0.6. 3. *URL: https://doi. org/10.5281/zenodo*, 2564164, 2019.

Shariq A Mobin. *Computational Models for Auditory Scene Analysis*. University of California, Berkeley, 2019.

Meinard Müller. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.

Adolfo G Navarro-Sigüenza and A Townsend Peterson. An alternative species taxonomy of the birds of mexico. *Biota Neotropica*, 4:1–32, 2004.

nti audio. Fast fourier transformation fft. `https://www.nti-audio.com/en/support/know-how/fast-fourier-transform-fft`. (Accessed on 14/12/2021).

Regina L Nuzzo. The box plots alternative for visualizing quantitative data. *PM&R*, 8(3):268–272, 2016.

Karan J Odom, Michelle L Hall, Katharina Riebel, Kevin E Omland, and Naomi E Langmore. Female song is widespread and ancestral in songbirds. *Nature Communications*, 5(1):1–6, 2014.

B Osgood. The fourier transform and its applications: Stanford university. *Stanford, California*, 94305:428, 2007.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011a.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011b.

Tereza Petrusková, Lucie Diblíková, Pavel Pipek, Eckehard Frauendorf, Petr Procházka, and Adam Petrusek. A review of the distribution of yellowhammer (emberiza citrinella) dialects in europe reveals the lack of a clear macrogeographic pattern. *Journal of Ornithology*, 156(1):263–273, 2015.

Pavel Pipek, Tereza Petrusková, Adam Petrusek, Lucie Diblíková, Mark A Eaton, and Petr Pyšek. Dialects of an invasive songbird are preserved in its invaded but not native source range. *Ecography*, 41(2):245–254, 2018.

Ilyas Potamitis, Stavros Ntalampiras, Olaf Jahn, and Klaus Riede. Automatic bird sound detection in long real-field recordings: Applications and tools. *Applied Acoustics*, 80:1–9, 2014.

Nirosha Priyadarshani, Stephen Marsland, Isabel Castro, and Amal Punchihewa. Birdsong denoising using wavelets. *PloS one*, 11(1):e0146790, 2016.

Nirosha Priyadarshani, Stephen Marsland, and Isabel Castro. Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology*, 49(5):jav–01447, 2018.

Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219, 2019.

Hannah Ritchie and Max Roser. Biodiversity. *Our World in Data*, 2021. https://ourworldindata.org/biodiversity.

Kenneth V Rosenberg, Adriaan M Dokter, Peter J Blancher, John R Sauer, Adam C Smith, Paul A Smith, Jessica C Stanton, Arvind Panjabi, Laura Helft, Michael Parr, et al. Decline of the north american avifauna. *Science*, 366(6461):120–124, 2019.

Zachary J Ruff, Damon B Lesmeister, Leila S Duchac, Bharath K Padmaraju, and Christopher M Sullivan. Automated identification of avian vocalizations with deep convolutional neural networks. *Remote Sensing in Ecology and Conservation*, 6(1):79–92, 2020.

Justin Salamon, Juan Pablo Bello, Andrew Farnsworth, and Steve Kelling. Fusing shallow and deep learning for bioacoustic bird species classification. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 141–145. IEEE, 2017.

Julia Shonfield and Erin Bayne. Autonomous recording units in avian ecological research: current use and future applications. *Avian Conservation and Ecology*, 12 (1), 2017.

Hans Slabbekoorn and Thomas B Smith. Bird song, ecology and speciation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357 (1420):493–503, 2002.

Julius O. Smith. *Spectral Audio Signal Processing*. http://ccrma.stanford.edu/ jos/sasp/, accessed <2019>. online book, 2011 edition.

Julius Orion Smith. *Mathematics of the discrete Fourier transform (DFT): with audio applications*. Julius Smith, 2007.

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

J Roberto Sosa-López and Daniel J Mennill. Continent-wide patterns of divergence in acoustic and morphological traits in the house wren species complex. *The Auk: Ornithological Advances*, 131(1):41–54, 2014.

Elias Sprengel, Martin Jaggi, Yannic Kilcher, and Thomas Hofmann. Audio based bird species identification using deep learning techniques. Technical report, 2016.

Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

Dan Stowell and Mark D Plumbley. Birdsong and c4dm: A survey of uk birdsong and machine recognition for music researchers. *Centre for Digital Music, Queen Mary University of London, Tech. Rep. C4DM-TR-09-12*, 2010.

Dan Stowell and Mark D Plumbley. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, 2:e488, 2014.

Mohammed Sunasra. Performance metrics for classification problems in machine learning, Feb 2019. URL `shorturl.at/hjpx8`.

Kyle A Swiston and Daniel J Mennill. Comparison of manual and automated methods for identifying target sounds in audio recordings of pileated, pale-billed, and putative ivory-billed woodpeckers. *Journal of Field Ornithology*, 80(1):42–50, 2009.

David Varodayan. Sampling and quantization. `https://courses.engr.illinois.edu/ece110/sp2021/content/courseNotes/files/?samplingAndQuantization`, 2014. (Accessed on 11/12/2021).

Peter Vary and Rainer Martin. *Digital speech transmission: Enhancement, coding and error concealment*. John Wiley & Sons, 2006.

Tuomas Virtanen, Mark D Plumbley, and Dan Ellis. *Computational analysis of sound scenes and events*. Springer, 2018.

Connor M Wood, Ralph J Gutiérrez, and M Zachariah Peery. The scientific naturalist. *Ecology*, 100(9):e02764, 2019.

Jie Xie and Mingying Zhu. Handcrafted features and late fusion with deep learning for bird sound classification. *Ecological Informatics*, 52:74–81, 2019.

Xiaoxia Zhang and Ying Li. Adaptive energy detection for bird sound detection in complex environments. *Neurocomputing*, 155:108–116, 2015.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1 (1-4):43–52, 2010.

List of Figures

List of Tables

List of Acronyms

DFT discrete Fourier Transform 7

STFT short-time Fourier transform 9

ANN Artificial Neural Network 64

ANNs Artificial Neural Networks 12

MFCCs Mel-Frequency Cepstral Coefficients 16

DCT discrete Cosine Transform 16

ROC Receiver Operating Characteristics 20

AUC Area Under ROC Curve 20

mAP Mean Average Precision 20

SVM Support Vector Machines 28

GMM Gaussian Mixture Model 29

HMM Hidden Markov Model 29

SNR signal-to-noise ratio 29

MLP multilayer perceptron 29

k-NN k-nearest neighbors algorithm 29

SMO Sequential Minimal Optimization 29

MARSYAS Music Analysis Retrieval and Synthesis for Audio Signals 29

ResNet Residual Neural Network 30