



Project Title

Optional Subtitle

Candidate Number: YPDR0¹

MSc Data Science & Machine Learning

Supervisor: Dr. Ifat Yasin

Submission date: 11 September 2023

¹**Disclaimer:** This report is submitted as part requirement for the MSc Data Science & Machine Learning at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged

Abstract

Summarise your report concisely.

Contents

1	Introduction	2
2	Background & related work	4
2.1	Evaluation metrics	4
2.2	Feature extraction	5
2.2.1	Features inspired by the human auditory system	7
2.2.2	Feature stacking	8
2.3	Classification	8
2.3.1	KNN	8
2.3.2	Decision trees	9
2.3.3	Gaussian Mixture Models	9
2.3.4	Hidden Markov Models	9
2.3.5	SVMs	9
2.3.6	Neural Networks	9
3	Methodology	10
4	Extensions of methodology	11
5	Conclusion	12

Chapter 1

Introduction

Birds play a crucial role in ecosystems around the world. They form an important link in the food chain, pollinate plants, and even plant trees [2]. The quantity and diversity of birds observed in an area can therefore be seen as a key indicator of the strength of its ecosystem.

Aside from being important ecological agents, birds also colour our lives with their sights, sounds and behaviours. The community of birdwatchers, known colloquially in the U.K. as ‘birders’ or ‘twitchers’, has grown considerably over the past few years. The Royal Society for the Protection of Birds (RSPB) reported a 70% increase in their website views over the first lockdown, with more than 50% of those views on pages looking at bird identification. The Bird Bird Garden Watch, an annual event which encourages people to note bird sightings in their own residences, brought in 1 million participants in January 2021, more than double the previous year’s tally. As such there is a growing commercial demand for accurate and easily accessible bird identification tools.

Birds are typically shy and protective creatures and tend to reside out of harm’s way in shrubs, trees and nests, and therefore are usually heard but not seen. The majority of their communication is done through distinctive vocalizations, known as birdsong, that are usually unique to an individual species. The consequence of this is that birds are typically identified through their birdsong rather than their visual sighting. Birdsong used for identification is usually recorded on microphones that may be running for a long time and/or located in a place that isn’t necessarily optimum to capture the vocalization. The recordings may be corrupted from a wide range of sources, such as ambient background noise, large changes in birdsong amplitude, long periods of silence, and vocalizations from other birds. Recordings captured by long running microphones may be several hours in

duration so it would be impractical to have a human expert identify the birds manually. There is therefore a need for robust birdsong identification tools that can handle these corruptions and that require minimal human intervention in order to classify unknown birdsong.

As with most machine learning classification problems, the work boils down to two key components. The first is feature extraction, that is, given an input signal usually in the form of an amplitude varying over time, how can it be transformed to a vector or matrix representation that captures the key information of the signal. The second is classification, i.e. given this representation in the feature space, how can it be used to train a model that can then perform classification on unseen samples of birdsong. There are further steps that can be taken to improve this process, such as pre-processing the signal in order to remove or reduce the influence of background noise [16].

There has been numerous research into audio classification problems both in birdsong and in other types of audio, such as speaker identification from human speech [9]. However, there remain novel methods that have been tested and have been shown to have good results in wider audio classification problems that have yet to be tried out in birdsong identification problems. These methods include both feature extraction and classification. The broad aim of this thesis is to attempt to evaluate these methods. In more detail, the aims of this thesis can be summarised as follows:

1. Explore the birdsong classification performance of feature representations shown to have promising results for non-birdsong related audio classification problems using simple classification models such as Support Vector Machines (SVM). The hypothesis is that feature representations shown to have good results in audio classification problems will have good results for birdsong identification.
2. Explore some of the hyperparameters available during the feature extraction process. The hypothesis is that using hyperparameter values more suited to birdsong classification problems will yield a higher classification accuracy.
3. Explore the performance of deep learning architectures such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) and compare the results with simpler statistical models. The hypothesis is that more complex and flexible architectures such as these will have superior performance when compared to simpler statistical models, such as SVMs.

Chapter 2

Background & related work

Describe here work that is connected to your thesis. This should include references to published work. There is no fixed rule, but I would expect a student to have read around 50 published research papers and reference them in a thesis.

2.1 Evaluation metrics

For multiclass classification problems, e.g. identifying a sample as a particular bird species from a list of $n > 2$ species, a simple classification accuracy is often used as an evaluation metric ([4], [17]). This is calculated as the percentage of correctly identified samples from a set of labelled samples previously unseen by the model.

Acevedo et al [1] used true positive (TP) and false positive (FP) rates to evaluate their models used to classify bird and amphibian calls. This evaluation method is sometimes preferred when analysing long recordings which may include several different species to be classified. TP gives an indication as to how well the model correctly identifies species present in the recording. FP indicates how often the model erroneously identifies species absent in the recording. A high performing model will have high values for TP and low values for FP .

Potamitis et al [16] used an alternative form of evaluation presented in terms of precision (P) and recall (R) which are defined as

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN} \quad (2.1)$$

where FN is the number of false negatives. Informally, P and R can be thought of as

$$P = \frac{\text{relevant retrieved instances}}{\text{all **retrieved** instances}}, \quad R = \frac{\text{relevant retrieved instances}}{\text{all **relevant** instances}} \quad (2.2)$$

It's well known that there is usually a trade-off between P and R , so usually the F -score is reported along with the P and R metrics. The F -score is defined as

$$F = \frac{(1 + \beta^2)PR}{\beta^2 P + R} \quad (2.3)$$


for some $\beta \in \mathbb{R}$.

For binary classification problems, i.e. classifying a sample as one of two classes (bird species), the Area Under Curve (AUC) metric is sometimes preferred [10]. The AUC ranges from 0 to 1, with a higher value indicating a better performing model. The AUC is sometimes desirable because it is scale-invariant (it measures how well predictions are ranked, rather than their absolute values) and classification-threshold-invariant (it measures the quality of the model's predictions regardless of what classification threshold is chosen.)

2.2 Feature extraction

As mentioned in Chapter 1, birds use vocalizations as a way to communicate with others. Birds have evolved over many thousands of years to make this form of communication very efficient in that it can travel long distances and cut through local ambient noise frequencies so that it can be received by other birds clearly. Some birds have even adapted their vocalizations so that they can be heard in local environments that have rapidly changed over the past decades, such as urban areas with increasing anthropogenic noise [12].

Bird vocalizations can be broadly divided into two main categories, calls and songs. Calls are typically short vocalizations that carry some specific function, such as warning others to the presence of a predator or calling others to flight [13]. Songs are usually longer and more acoustically complex and occur more spontaneously. Songs are typically employed as breeding calls or territorial defence. While all birds produce calls, in many species of birds only the males utilise songs and often only during breeding season. The song and call of a Eurasian Wren can be seen and contrasted in figure 2.1.

Doupe et al [6] showed that there were striking similarities between birdsong and human language. Similar to human language, birdsong can be thought of as comprised of hierarchical levels of phrases, syllables, and elements [3].  Raw recordings

of birdsong often contain periods of silence that occur between phrases which are unlikely to provide useful information in order to train a machine learning model. Fagerlund [8] showed that a good level of accuracy can be achieved by training models on features extracted from segmented syllables which were used as training samples. An algorithm for robust syllable segmentation was proposed in [7] and is used in this thesis. The algorithm is described in more detail in Section.

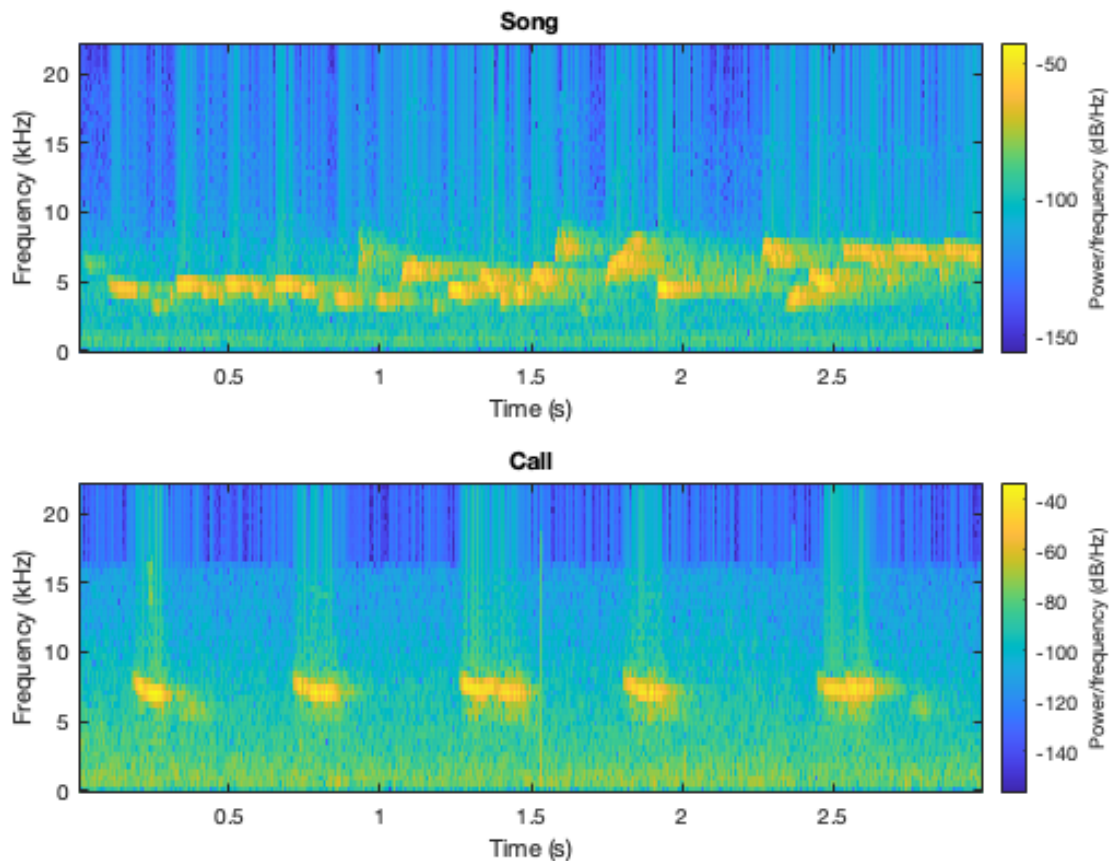


Figure 2.1: Spectrograms of the call and song of the Eurasian Wren (*Troglodytes troglodytes*). As can be seen, the song is much more complex in terms of pitch and acoustic structure compared to the call

Once the syllables have been segmented, there exists an abundance of options to turn the raw syllable input signal into features that might be used as training samples. In the following sections some of the more popular feature representations are described, along with some novel methods that have yet to be tried on birdsong. Note that the following list is certainly not exhaustive and emphasis has been placed on feature representations

that are relevant to this thesis.

2.2.1 Features inspired by the human auditory system

Research has shown that birdsong and human language share many similarities in terms of the neural mechanisms employed to form the language/song, the impact of social contact in learning the language/song, and so on [6]. Given that the human auditory system has evolved over thousands of years to best process human speech, it seems reasonable to suggest that using features based on the human auditory system may be effective when it comes to birdsong.

At a high level, the human auditory system works by translating changes in air pressure originating from a source and reaching the outer ear of a listener into vibrations that travel along an internal organ known as a cochlea. The vibrations trigger electric signals that move along auditory nerves to the brain, where they are interpreted as sounds. The physiology of the cochlea means that certain parts of the organ are more sensitive to certain frequencies of vibrations, so different frequencies will lead to different electrical signals moving to the brain. This allows the brain to learn to differentiate between frequencies.

Mel frequency cepstrum coefficients

The makeup of the human auditory system means that humans have a greater ability to differentiate pitch at lower frequencies than they do at higher frequencies. In other words, humans perceive pitch non-linearly. This has led to the development of a logarithmic scale, known as the mel scale, such that equal distances on the scale have the same *perceptual* distance.

The mel frequency cepstrum coefficients (MFCC) are part of a family of cepstrum coefficients that capture information about the rate of change in different spectrum bands. MFCC differs from other cepstrum coefficients in that it uses the mel scale to transform the spectrum of an input signal, thus utilising the human auditory system in the calculation of its coefficients.

The i -th mel cepstral coefficient is computed as [5]

$$\text{MFCC}_i = \sum_{k=1}^K X_k \cos \left(\frac{i(k - 0.5)\pi}{K} \right) \quad (2.4)$$

where X_k is the logarithmic energy of the k -th mel-spectrum band, and K is the total

number of the mel-spectrum bands. Usually 8–13 MFCC coefficients are used as the feature vector representing one time frame of the signal. The 0th coefficient is often excluded as it represents the average log energy of the signal and is unlikely to carry any relevant information to help with classification.

MFCCs are often presented with their delta (Δ) and double-delta ($\Delta\Delta$) values that capture the local temporal dynamics and temporal changes of the delta values respectively.

MFCCs are used as feature representations since they are simple to compute and have been shown to have good performance in a wide range of audio classification tasks, such as speaker identification [14] and emotion recognition [11]. MFCCs have also been shown to lead to good classification accuracy for birdsong identification problems ([8] and [17]).

`\include graph of filterbank`

Gammatone cepstrum coefficients

Gammatone cepstrum coefficients (GTCCs) are similar to MFCCs except their calculation uses gammatone filterbanks instead of mel scale filterbanks. Gammatone filterbanks are designed to simulate the motion of the membrane inside the cochlear, known as the basilar membrane, when it is exposed to vibrations transmitted by the outer and middle ear [15].

A gammatone filter with a centre frequency f_c is defined as

$$g(t) = at^{n-1}e^{-2\pi bt} \cos(2\pi f_c t + \psi) \quad (2.5)$$

where t is the time in seconds, ψ is the phase in radians (usually set to 0), $a \in \mathbb{R}$ controls the gain, n is the filter's order and b is the filter's bandwidth in Hz.

Similar to MFCCs, GTCCs are typically used with their Δ and $\Delta\Delta$ values and have also been shown to have a good performance in non-speech audio classification problems [19].

`\include graph of filterbank`

Multi resolution cochleagram

2.2.2 Feature stacking

2.3 Classification

2.3.1 KNN

KNN used by [18]

2.3.2 Decision trees

Description of decision trees. Used by [1].

2.3.3 Gaussian Mixture Models

2.3.4 Hidden Markov Models

2.3.5 SVMs

2.3.6 Neural Networks

Feedforward Neural Network

Convolutional Neural Network

Recurrent Neural Network

Chapter 3

Methodology

Describe your method in detail and with great clarity, distinguishing it from other works (if it is indeed a novel idea). It is very important to clearly motivate your method.

Describe the results of your method here in this chapter.

Chapter 4

Extensions of methodology

It is unlikely that everything you tried worked well, so in this chapter you may wish to describe a modified version of your method and the associated results. Explain why you were motivated to try this extension and how you think it might help to address some of the shortcomings of the method in Chapter 3.

Chapter 5

Conclusion

Summarise what you have achieved and evaluate honestly if you feel the approach has been largely successful. Explain what could be improved still and perhaps why the method is not working well (if that is the case).

Bibliography

- [1] Miguel A Acevedo, Carlos J Corrada-Bravo, Héctor Corrada-Bravo, Luis J Villanueva-Rivera, and T Mitchell Aide. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, 4(4):206–214, 2009.
- [2] Richard K Broughton, James M Bullock, Charles George, Ross A Hill, Shelley A Hinsley, Marta Maziarz, Markus Melin, J Owen Mountford, Tim H Sparks, and Richard F Pywell. Long-term woodland restoration on lowland farmland through passive rewilding. *PloS one*, 16(6):e0252466, 2021.
- [3] Clive K Catchpole and Peter JB Slater. *Bird song: biological themes and variations*. Cambridge university press, 2003.
- [4] Deep Chakraborty, Paawan Mukker, Padmanabhan Rajan, and Aroor Dinesh Dileep. Bird call identification using dynamic kernel based support vector machines and deep neural networks. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 280–285. IEEE, 2016.
- [5] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.
- [6] Allison J. Doupe and Patricia K. Kuhl. Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience*, 22(1):567–631, 1999. PMID: 10202549.
- [7] Seppo Fagerlund. Automatic recognition of bird species by their sounds. *Finlandia: Helsinki University Of Technology*, 2004.

- [8] Seppo Fagerlund. Bird species recognition using support vector machines. *EURASIP Journal on Advances in Signal Processing*, 2007:1–8, 2007.
- [9] Mineichi Kudo, Jun Toyama, and Masaru Shimbo. Multidimensional curve classification using passing-through regions. *Pattern Recognition Letters*, 20(11):1103–1111, 1999.
- [10] Yi Ren Leng and Huy Dat Tran. Multi-label bird classification using an ensemble classifier with simple features. In *Signal and information processing association annual summit and conference (APSIPA), 2014 Asia-Pacific*, pages 1–5. IEEE, 2014.
- [11] MS Likitha, Sri Raksha R Gupta, K Hasitha, and A Upendra Raju. Speech based human emotion recognition using mfcc. In *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*, pages 2257–2260. IEEE, 2017.
- [12] David Luther and Luis Baptista. Urban noise and the cultural evolution of bird songs. *Proceedings of the Royal Society B: Biological Sciences*, 277(1680):469–473, 2010.
- [13] PETER MARLER. Chapter 5 - bird calls: a cornucopia for communication. In Peter Marler and Hans Slabbekoorn, editors, *Nature’s Music*, pages 132–177. Academic Press, San Diego, 2004.
- [14] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010.
- [15] Roy D Patterson, KEN Robinson, John Holdsworth, Denis McKeown, C Zhang, and Michael Allerhand. Complex sounds and auditory images. In *Auditory physiology and perception*, pages 429–446. Elsevier, 1992.
- [16] 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.
- [17] Murugaiya Ramashini, Pg Emeroylariffion Abas, Kusuma Mohanchandra, and Liyanage C De Silva. Robust cepstral feature for bird sound classification. *International Journal of Electrical and Computer Engineering*, 12(2):1477, 2022.

- [18] Murugiaya Ramashini, Pg Emeroylariffion Abas, Ulmar Grafe, and Liyanage C De Silva. Bird sounds classification using linear discriminant analysis. In *2019 4th International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–6. IEEE, 2019.
- [19] Xavier Valero and Francesc Alias. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE transactions on multimedia*, 14(6):1684–1689, 2012.