WAVELET-BASED BIRDSONG RECOGNITION FOR CONSERVATION

A thesis presented in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
in
Computer Science
at Massey University, Palmerston North,
New Zealand.

Nirosha Priyadarshani

2017

To Sumudu, for everything, forever...

Contents

List of Tables

List of Figures

Abstract

According to the International Union for the Conservation of Nature Red Data List nearly a quarter of the world's bird species are either threatened or at risk of extinction. To be able to protect endangered species, we need accurate survey methods that reliably estimate numbers and hence population trends. Acoustic monitoring is the most commonly-used method to survey birds, particularly cryptic and nocturnal species, not least because it is non-invasive, unbiased, and relatively time-effective. Unfortunately, the resulting data still have to be analysed manually. The current practice, manual spectrogram reading, is tedious, prone to bias due to observer variations, and not reproducible.

While there is a large literature on automatic recognition of targeted recordings of small numbers of species, automatic analysis of long field recordings has not been well studied to date. This thesis considers this problem in detail, presenting experiments demonstrating the true efficacy of recorders in natural environments under different conditions, and then working to reduce the noise present in the recording, as well as to segment and recognise a range of New Zealand native bird species.

The primary issues with field recordings are that the birds are at variable distances from the recorder, that the recordings are corrupted by many different forms of noise, that the environment affects the quality of the recorded sound, and that birdsong is often relatively rare within a recording. Thus, methods of dealing with faint calls, denoising, and effective segmentation are all needed before individual species can be recognised reliably. Experiments presented in this thesis demonstrate clearly the effects of distance and environment on recorded calls. Some of these results are unsurprising, for example an inverse square relationship with distance is largely true. Perhaps more surprising is that the height from which a call is transmitted has a significant effect on the recorded sound. Statistical analyses of the experiments, which demonstrate many significant environmental and sound factors, are presented.

Regardless of these factors, the recordings have noise present, and removing this noise is helpful for reliable recognition. A method for denoising based on the wavelet packet decomposition is presented and demonstrated to significantly improve the quality of recordings. Following this, wavelets were also used to implement a call detection

algorithm that identifies regions of the recording with calls from a target bird species. This algorithm is validated using four New Zealand native species namely Australasian bittern (*Botaurus poiciloptilus*), brown kiwi (*Apteryx mantelli*), morepork (*Ninox novaeseelandiae*), and kakapo (*Strigops habroptilus*), but could be used for any species. The results demonstrate high recall rates and tolerate false positives when compared to human experts.

Acknowledgements

Undertaking this PhD has been a life-changing experience for me, inspired by the support and encouragement from researchers, conservation groups, Department of Conservation (DOC), and interesting individuals, combined with the best supervisory panel one could expect to have. Thinking of all the help I received reminds me of how fortunate I am, it would not have been possible to do this project without the support and guidance I received from those people.

I would like to express my sincere gratitude to Prof. Stephen Marsland and A/Prof. Isabel Castro for your effective supervision, understanding, patience, encouragement, criticisms, and for always being there to offer advice. Working with you was enjoyable as you lived in the project and enjoyed it. Thank you very much for promptly reading the manuscripts and providing valuable comments to improve my writing and my work in general. Stephen, thank you for your eternal and active support, you always had a solution when I struggled with a problem, it was sometimes challenging to achieve, but you always guided me to think outside the box. Most importantly, you were able to pull me back to the right path (even without me knowing) when I was lost. Thank you for independently implementing most of the methods we tried helping me to boost my confidence and also to fix bugs! I have learned a lot from you and appreciate your simple explanations given to complex concepts. I thank you for writing that machine learning book and allowing me to follow most of its content through your lectures during the first year of this study although our initial machine learning experiments are not in the thesis. The Latex template of Massey thesis that I used to create this document made my life easier during the last few months of the study. I also appreciate your financial support that helped me to focus on studies. Overall, I was blessed to have a supervisor like you. Isabel, thank you for introducing me to the avian world and to the field of conservation, your input, time, and energy into the project has been invaluable. I do believe that you were the main force to shape the project towards conservation opening a fairly large network of ornithologists, bird conservationists, and community groups. Thank you for guiding me in the field, particularly to handle the manual recorders in the field and to carefully get close to the birds for close-range recordings. It was certainly a fun time that I spent in the field with you. Thanks you so much for being

so generous with your time and your hours spent on spectrogram readings. Thank you for encouraging me to face the challenges, being there with me, and giving me moral support as a supervisor as well as a friend.

I owe a big thank to Dr. Amal Punchihewa, co-supervisor, for accepting me as a PhD student in first place, initiating this project, helping me all the way to Massey. It was you who gave me the initial idea of this research project. Subsequently, you helped me to develop the idea and design this project. Although when I arrive in New Zealand you transferred your role as the main supervisor to Stephen and moved overseas, you continued to support me remotely by helping me in various ways. You have been reachable throughout the project despite your tight schedule and time difference between the countries. You also helped me to improve my presentation skills by listening to my trial presentations via Skype and making constructive comments. I am really grateful to you for your all encouragement, kind advice and guidance throughout the project. I must also thank the referees who reviewed the journal article, as well as to the conference attendees who posed important questions and suggestions related to my research.

I am grateful to the volunteers who turned up with a short notice and helped me in the field with playback re-capture experiments, Natasha, Tim, Sumudu, Catherine, Kim, Julia, Ross Bell, Myung Jong, Anindya, Asif, Tobby, Emma, Giulian, Janna, Kelly, and Shari. Again thank you Natasha, Tim, Sumudu, and Catherine for your ability to be positive even when lots of supplejacks are ahead and need to be passed to make a straight line of recorders! Thank you Sara Treadgold (DOC, Whanganui branch), for lending me acoustic recorders, Dave Bell (Coordinator, NZ Falcon Survey) for connecting me to the right people (and also for donating us a Song Meter recorder with all the accessories), and Emma Williams for lending me few more acoustic recorders. Thank you, Paul Barrett and Cleland Wallace for your great technical support, for making the essential setups in-house. A big thank to Gary Mack (Massey Grounds Manager) not only for facilitating my use of the rugby field for the playback experiments but also to the kind efforts taken to minimise the machine noise by pausing the maintenance work during the trials. I am thankful to James Lambie (Science Coordinator, Horizons Regional Council) who provided access to the Pohangina Reserve and all the details about the tracks and safety matters in the forest. I would like to thank Ellen Schooner (Ecology Group) for connecting me to James and also for facilitating me to join with the Mokoia Island field visit. Without all your help Chapter 3 would simply not exist in this thesis.

Thank you Emma Williams, our bittern consultant, for welcoming me to the Hatuma lake and facilitating me (and my family) to spend several nights there and feel/-hear the amazing bittern booms; thanks for the inspirational moments you shared with

me. You helped me throughout the thesis providing time-stamped bittern recordings and opening new connections to the project. I saw your true dedication to save bitterns and you motivated me, which was so useful throughout this journey. To Rebecca, I was fascinated by your desperate work to save Samoan birds, particularly the Manumea (tooth-billed pigeon) and Ma'oma'o. They led me to test and confirm our methods on Manumea recordings, although this is not included in this thesis. I also thank Alex Brighten for giving me the opportunity to tune the selected commercial software to process a year-around continuous morepork recordings made in Ponui Island. Even though it took lots of my time, it was a great experience that convinced me of the difficulties associated with my research. Also, I highly appreciate the support given in the field during my visit to Ponui and providing me the whole morepork dataset to run my experiments. During the software tuning for morepork, Jeff Knewstubb gave me a great support as an experienced user and I appreciate that. I would like to thank Andrew Digby (Science Advisor for kākāpō and takahē, DOC) for your input to this project in several ways. Your nicely organised little spotted kiwi recordings (both manual and autonomous) and kākāpō recordings (collaborating with Bruce C. Robertson, University of Otago) were very useful for this thesis. I appreciate your suggestions and also networking me with other potential groups within DOC. I am also thankful to Les McPherson who kindly shared some birdsong from his Natural History Unit Sound Archive (http://www.archivebirdsnz.com/) with me at the beginning of the study.

I thank my home university, University of Kelaniya, for offering me an ideal environment in which I felt free to concentrate on this research. I acknowledge the funding received from Higher Education for the Twenty-first Century project (HETC), Sri Lanka. I am also grateful to the funding received from SEAT, Massey University. Financial support and the Conservation Innovation Award 2014 received from WWF facilitated and encouraged me to complete this thesis. I also appreciate the recent research fund granted from J S Watson Trust 2016, Forest & Bird to continue software implementations extending this thesis, particularly for kiwi conservation involving with community groups. I thank Richard Witehira (Blandy) and Jason Taiaroa for helping to setup the community groups for the latter. Thanks to Dilantha Punchihewa, Michele Wagner, Linda Lowe, and Karen Pickering for all the administrative support throughout the research.

I thank my office-mates and friends, Pramila, Ishani, Nadee, and Nisansala. Last, but not least, I am very grateful to my family – parents for their support, upbringing, and believing in me – loving husband Sumudu and our little princess, Dinara, for all your dedication during this research. Sumudu not only took the responsibilities at home providing a perfect environment for me, but also proofread the entire thesis. Dinara, you made me happy during the hard times – I love your surprise picnics!

Chapter 1

Introduction

This chapter provides an overview of the research background and identifies the scope of the thesis. It presents the aims and objectives and briefly summarises the structure of the thesis and an outline of the rest of the chapters.

1.1 Motivation

Biodiversity loss has become a major global environmental issue, principally due to human expansion activities, which have meant habitat destruction, over-exploitation and unsustainable use of resources, and global climate change. Nearly a quarter of extant bird species are threatened according to the International Union for the Conservation of Nature Red Data List (IUCN, 2014); 1,373 (more than 13%) of the total world bird species are vulnerable or in immediate danger of extinction. Taking New Zealand as an example, our avifauna evolved in the absence of mammals and therefore contains some unique phylogenetic groups and unusual adaptations. After the arrival of humans 45% of species became extinct due to habitat destruction, the introduction of mammals, and disease – 50% of remnant species are currently threatened by extinction from the same causes (Tennyson and Martinson, 2006; Miskelly et al., 2008; Carolyn et al., 2015; Department of Conservation).

One challenge for managing bird populations is knowing how many individuals are present in a region. Reliable methods are needed to measure population sizes and population trends of bird species in order to evaluate their conservation status and to implement wildlife management programs effectively. Conventional observer-based survey methods such as the five minute bird count used by New Zealand Department of Conservation (performed by trained field observers) and similar methods (Barraclough, 2000; Taylor and Pollard, 2008) are costly and can only cover very limited spatial and temporal regions. This lack of systematic, cost-effective, and multi-scale survey methods is a major obstacle to protecting the extant avifauna.

1.2 The Problem

During the last decade, there has been significant progress in the field of bio-acoustics, particularly in hardware development. Nowadays, there are affordable autonomous recorders that can be programmed to record at specific times. These recorders have storage capacity and can use energy sources (solar/batteries) that allow them to be deployed for long periods. The recordings can be transmitted to the laboratory real-time using wireless sensors or can be collected as a batch by returning to the field weeks or months after their deployment (Wimmer et al., 2013; Stattner et al., 2012). Therefore, conservation managers increasingly use these recorders to infer presence/absence and attempt to measure the abundance of birds.

Compared to the hardware revolution, the software for automatic processing of recordings – machine recognition of birdsong from field recordings – is not well developed. The current practice, vetting the recordings by spectrogram reading and/or listening by experienced observers, is not feasible for the processing of the large volumes of recordings that accumulate from multiple recorders and multiple sites (Taylor, 1995; Brighten, 2015; Wimmer et al., 2013; Colbourne and Digby, 2016).

Although there is large amount of literature available for automated birdsong recognition (with methods mostly based on human speech recognition), either they are focussed on general use, not conservation (which requires high accuracy) or they are extremely species-specific, requiring new methods for each new species. Even when they are intended for conservation purposes, the majority of researchers use carefully selected (often manually recorded) high-quality recordings to assess the methods, making them likely to fail in the real world, where autonomous recorders are used. Accordingly, there is a dire need for a robust, scalable, and user-friendly automatic method to analyse unattended long duration field recordings. This thesis works towards this aim, and identifies the outstanding issues needed to find a practical automated solution.

1.3 Aims and Objectives

The aims of the thesis are:

1. to study the challenges associated with the problem of birdsong recognition from automatic field recordings, such as the effects of noise;

2. to contribute to the development of methods to automate the birdsong recognition process.

The objectives set to achieve the aims are:

1. to investigate how environmental factors and distance to the recorder affect the recording of birdsong;

2. to develop a method to remove noise from recordings;

3. to develop a method to automatically detect calls from any target species from continuous field recordings collected with autonomous recorders with high accuracy.

1.4 Thesis Outline and Research Contributions

The thesis was written as a collection of four independent papers. One has already been published and the others will be submitted soon after the completion of the thesis (for a list, see Table 1.1). I am the principal author on each of these papers, with my supervisors as co-authors. Chapters follow the style of the journals with minor modifications to the formatting. Since it is a collection of papers, there is unavoidably some repetition between the chapters of the thesis, particularly in the introductions. Scientific names of species are provided upon the first use in each chapter.

The literature review (Chapter 2) is based upon all of the papers that I could find that study automatic recognition of birdsong, and provides a general introduction to the relevant topics (signal processing, machine learning, surveying methods, etc.). The main contribution of the chapter is a series of tables that summarises the relevant papers for each of the particular parts of a birdsong recogniser: pre-processing and noise reduction; segmentation; feature extraction; classification. My principal conclusion is that while there is a large amount of machine learning work that uses small datasets of birdsong as a demonstration system, there are relatively few large-scale studies of the depth necessary to make birdsong recognition useful for ecologists, wildlife managers, and community groups.

One key gap in the literature that I identified was the lack of study of precisely how the quality of recordings degraded with distance between the calling bird and the recorder, and how environmental factors modified this. Chapter 3 presents a playback and re-capture experiment designed to fill this gap. Twenty automatic recorders made by the New Zealand Department of Conservation, and widely available across New Zealand, were used to record broadcasts of a set of clean birdsong from eleven species native to New Zealand. The signal-to-noise ratio for each recording of birdsong was calculated at each recorder. We used an area of forest and an open grassland area in different environmental conditions, placing the recorders at a variety of distances and directions from the speaker. We evaluated the predictive value of a set of environmental and recording-specific factors using a generalised linear model. The results revealed the factors that significantly affect the quality of birdsong acquisition: this can be used in order to establish protocols for deploying the recorders in the field for recording species with particular types of call.

Following the results from Chapter 3, it was clear that in order to be successful in automated recognition, there is a need for removing noise from recordings, particularly background environmental noise. I chose to use wavelets for this as they provide a solution to the trade-off between time and frequency resolution that is inherent in conventional Fourier analysis. Chapter 4 presents an algorithm that improves the signal-to-noise ratio of natural noisy field recordings by an order of magnitude and demonstrates it on a variety of recordings from different species of birds.

The wavelet decomposition of the sounds can also be used as features for classification of birdsong. In general, it is necessary to perform segmentation of the calls from the sound file before using machine learning or other techniques to recognise them. However, the wavelet approach gives us an interesting way to perform both simultaneously: I seek elements of the wavelet packet tree that correlate well with calls from particular species, and segment only those calls in the file. I call this *species-specific segmentation by wavelet filtering* since only the calls corresponding to the species of interest are considered. The aim is to reliably identify calls from one particular species, which can be used for presence/absence studies and also to estimate call counts. Chapter 5 presents the algorithm of this method and assesses it on unattended field recordings of the Australasian bittern (*Botaurus poiciloptilus*), brown kiwi (*Apteryx mantelli*), morepork (ruru; *Ninox novaeseelandiae*), and kākāpō (night parrot; *Strigops habroptilus*). As a comparison with other methods, we also use a third-party dataset of the American robin (*Turdus migratorius*). The results demonstrate very high recall when detecting (close-range) loud (>95%) and even very faded (approximately 70%) birdsong.

In this thesis I have performed the initial work required to be able to satisfactorily identify calls of birds and count them at the level required for bird conservation. However, there is rather more work to do, and I discuss this in relation to the research in the field in Chapter 6, where I suggest future work that is required if automatic birdsong recording and recognition is to be useful to people monitoring wild bird populations.

Table 1.1: Manuscripts published or in preparation for submission from material in this thesis.

Chapter	Title	Journal	Status
2	Automated Birdsong Recognition in Complex Acoustic Environments: A Review	To be decided	In prep.
3	The Impact of Environmental Factors in Birdsong Acquisition using Automated Recorders	To be decided	In prep.
4	Birdsong Denoising using Wavelets	PLOS One	Published, Jan 2016
5	Birdsong Recognition in Continuous Field Recordings	Ecology and Evolution	In prep.

Chapter 2

Automated Birdsong Recognition in Complex Acoustic Environments: A Review

Abstract

Nearly a quarter of the world's avian species are either threatened or at risk of extinction. Hence, there is a great deal of attention placed on bird conservation, and this requires accurate population estimates. With the availability of programmable acoustic recorders, conservationists are increasingly using them to determine the presence, abundance and decline of various species. Unlike humans, these recorders can be left in the field for extensive periods of time, allowing them to capture all sounds produced in any habitat, including rare birdsong. Although data acquisition is automatic, manual processing of recordings is labour intensive, difficult and tedious, as well as being prone to bias due to observer variations. Automating the birdsong recognition process so that it can successfully process unattended field recordings is thus a desirable research aim: a successful tool will enable the real benefits of programmable acoustic recorders to be felt, and provide the data needed by conservation practitioners and decision makers to protect the extant avifauna.

Given the clear need, many researchers and companies have developed birdsong recognisers. However, most ecologists and conservationists do not utilize them to process natural field recordings because the system calibration time is exceptionally high and requires considerable knowledge in signal processing and underlying systems, making the tools complicated and less user-friendly. Even allowing for these difficulties, getting accurate results is exceedingly hard. The purpose of this paper is to review

the work that has been done to date in this area and propose possible future directions for further developments in automated birdsong recognition. We do this by first looking into the essentials of a birdsong recogniser, and then analysing the strengths, weaknesses, and challenges of the currently available methods.

2.1 Introduction

Approximately 10,000 bird species exist on earth; this number exceeds the mammals, reptiles and amphibian species (Unwin, 2011). Birds are widely spread, with different species occupying habitats from deserts, to mountain tops, valleys and even ice caps. According to the International Union for the Conservation of Nature Red Data List (IUCN, 2014) 1,373 (more than 13%) of the total world bird species are vulnerable or in immediate danger of extinction. Therefore a great deal of attention is placed on bird conservation worldwide.

Effective bird monitoring methods are needed to assess species abundance and evaluate the consequences of current species management-for-conservation practices. This is particularly important because the trends in bird populations provide an indication of species diversity, abundance and overall balance of a given biome (Towsey et al., 2012; Digby et al., 2013; Dawson and Efford, 2009; Vielliard, 2000). However, methods to accurately estimate bird population sizes require a great deal of time and effort and are costly and for these reasons are only applied at small scales (Sutherland et al., 2004). Conservation managers need cost-effective tools to monitor the changes in population size of the species they manage, often in difficult terrain and over large areas. Birdsong is often used to detect, monitor, and quantify species because it works even when the individuals are out of sight.

The common approach of estimating populations, the call count (point count) surveys – such as the five minute bird count method used in New Zealand (Department of Conservation) that is performed by trained field observers – are labour intensive and prone to bias, depending on the expertise and hearing capacity of individual observers (Brandes, 2008; Rosenstock et al., 2002; Sauer et al., 1994). Therefore, it is important to consider the auditive capacity and experience of surveyors when analysing such data (Emlen and DeJong, 1992). It is also suggested that the training of novice volunteers can improve the reliability of bird surveys (McLaren and Cadman, 1999). The results of point count call surveys are also subjective, have high errors when the call rate is high, and the presence of observers can affect the vocal activity of the birds (Bye et al., 2001). These surveys are usually done during fine weather in easily accessible areas, therefore they can be biased by weather conditions and location.

As call count surveys are short (usually 5-10 minutes; (Dawson and Bull, 1975; Angehr et al., 2002)), they cannot fully describe temporal patterns (Digby et al., 2013;

Potamitis et al., 2014): for example, Vielliard (2000) and Loyn (1985) observed that call counts over periods of less than 20 minutes underestimated rare species. Practical comparisons between long time recordings by autonomous recording units and human observers have confirmed that the former detects many more species (Cunningham et al., 2004).

Today, high-end weather-proof bio-acoustic recorders with long battery life and high memory capacity are available for affordable prices (e.g. the recorders made by New Zealand Department of Conservation and Song Meter recorders from Wildlife Acoustics Inc., Concord, MA, USA). These automatic recorders are specially designed for collecting long autonomous field recordings with minimum human intervention. One can schedule the recorders and mount them in the field and return weeks or months later for the data, or set up a sensor network to directly download data to the laboratory (Wimmer et al., 2013; Stattner et al., 2012). The recorders can be operated in 24/7 mode, meaning that they are capable of capturing both the diurnal and nocturnal sonic environment, including any rare or cryptic bird vocalisations, in any habitat, including ecologically sensitive areas or areas that are difficult to access. Accordingly, conservation managers are increasingly interested in using these unattended (automated) recordings to infer the presence, abundance and decline of their target species. After collecting the recordings, they are generally processed through spectrogram reading and/or listening by experienced observers, which remains a labour intensive task (Taylor, 1995; Brighten, 2015; Wimmer et al., 2013; Colbourne and Digby, 2016): it is not feasible to manually process weeks or months worth of field recordings. Thus, automated bird song recognition could play an important role in environmental monitoring if the recogniser is capable of processing noisy field recordings and producing robust results. While the development of such automated recognisers is on-going, the need for further developments is evident given that ecologists and conservation managers still spend a great deal of time manually scanning field recordings because none of the available automated birdsong recognition software fulfil their requirements reliably (Ulloa et al., 2016; Swiston and Mennill, 2009; Goyette et al., 2011; Potamitis, 2014; Potamitis et al., 2014). This lack of appropriate resource motivated us to review the current literature and to identify potential methods to fill in the gaps.

When dealing with field recordings as primary data, there are some common challenges regardless of the methods used to automate the process (see Box 2.1). Any automated method needs to be tailored to these variations. In order to develop a robust recogniser, it is generally essential to have a rich database including good data of the possible variations in vocalisations of each species. Obtaining these data is a challenging task that requires expertise in handling recorders, and cooperative animals (many birds tend to avoid humans). For rare birds these data can be exceedingly hard

to obtain.

Box 2.1: Challenges associated with implementation of automated birdsong recognition to process field recordings.

1. There is a plethora of unavoidable environmental noise overlapped with field recordings. Noise exists in lesser or greater degrees in all recordings, but those made in natural environments are subject to a variety of extraneous sounds, both biological and non-biological.

2. Bird vocalisations are of varying power. Even though the recorder is fixed (mounted) in the field, the birds can be anywhere, some closer and some further away, and at different angles to the recorder's microphones. Accordingly, some songs are louder and some are quieter in the recordings. During song detection (segmentation), normally the faint songs tend to not be included, but this can be inconsistent depending upon the noise level. The challenge is to maintain accuracy while improving the sensitivity to the target sounds.

3. Birds (of the same or different species) call on top of each other, for example when duetting, during the dawn or dusk chorus, and when they live in flocks.

4. There is large inter- and intra-species song variability. Birds maintain their own song repertoire, with the size of the repertoire and the complexity varying across the species. Some species repeat the same song, while others have a variety of songs and are capable of creating new songs. Some bird species, such as tui (*Prosthemadera novaeseelandiae*), exhibit geographical variations on their songs (Hill et al., 2013). Although this phenomenon is a challenge in species recognition, it could possibly allow individual recognition (Cheng et al., 2010; Ptacek et al., 2016; Baldo and Mennill, 2011; Dent and Molles, 2016; Gilbert et al., 1994).

5. Similarly to human speech, a bird may generate the same song with short or long duration in different situations. Further, birds are capable of adapting their sounds according to the environment.

6. Birds generate incomplete/quick calls in critical situations, especially during the breeding season when they are occupied with incubation and/or chick rearing.

7. During the song learning process, juveniles produce unusual calls, making the recognition more complicated (Williams, 2004) and sometimes, even human experts fail to recognise the species from hearing a juvenile.

In this paper we review (1) the published methods for the automatic processing and recognition of birdsong illustrating their strengths and weaknesses and (2) the current available birdsong recognition-related software, discussing their benefits and disadvantages. We close by making some conclusions, recommendations, and outlining

future directions of research.

2.2 Birdsong Processing and Recognition

Birds have an innate ability for processing and recognising species-specific vocalisations (Theunissen and Shaevitz, 2006; Cheng, 2008; Matsunaga and Okanoya, 2009; Zollinger and Brumm, 2015; Reichard and Anderson, 2015; Elemans, 2014; Wilbrecht and Nottebohm, 2003; Perkel and Farries, 2000; O'Loghlen and Beecher, 1999). Humans are capable of identifying birds both aurally and visually: the average person can recognise bird species in their backyard, while experts can identify thousands of bird species by their song alone. The underlying process involves listening (or viewing the spectrogram), remembering (developing a mental library), and generalising. Accordingly, automating this process involves several building blocks (Fig. 2.1). This process is usually aided by initial visual inspection of the birdsong, particularly spectrogram reading (see Section *Visual Representation of Birdsong*). In this section we discuss each constituent part of an automated birdsong recogniser, referring to the relevant literature, which is summarised in Tables 2.1–2.3.

2.2.1 Sound Recording and Storage

Calling birds produce slight fluctuations of air pressure, which the auditory systems can resolve as sound, as can a microphone; the latter turns these air vibrations into voltages so that it is possible to represent them digitally (Mindlin, 2013). To obtain a discrete time digital signal from a continuous time analogue signal, the continuous signal is sampled at equally spaced intervals. The *Nyquist sampling theorem* (Landau, 1967) stipulates the minimum sampling frequency required to represent a continuous time signal with a discrete time series. The theorem states that a signal can be recovered from its samples if the sampling frequency is at least twice the highest frequency of the original signal. The dynamic range of values provided to record the samples is determined by the resolution, the number of bits per sample. Larger resolutions in amplitude sampling bring larger dynamic ranges and more resolution, but occupy more space. Sound files accumulate rapidly when large numbers of automatic recording units are operated continuously, meaning that using an effective method for storing and retrieving sound inventory is crucial in long-term acoustic monitoring.

Sound recordings are needed initially to develop the birdsong database that is used as reference templates during the recognition; developing a labelled database covering all possible birdsong variations of each species is a challenging task. The size of the training data required mainly depends on the target species, with fewer data required for species with small repertoires and simpler vocalisations. Developing the training data

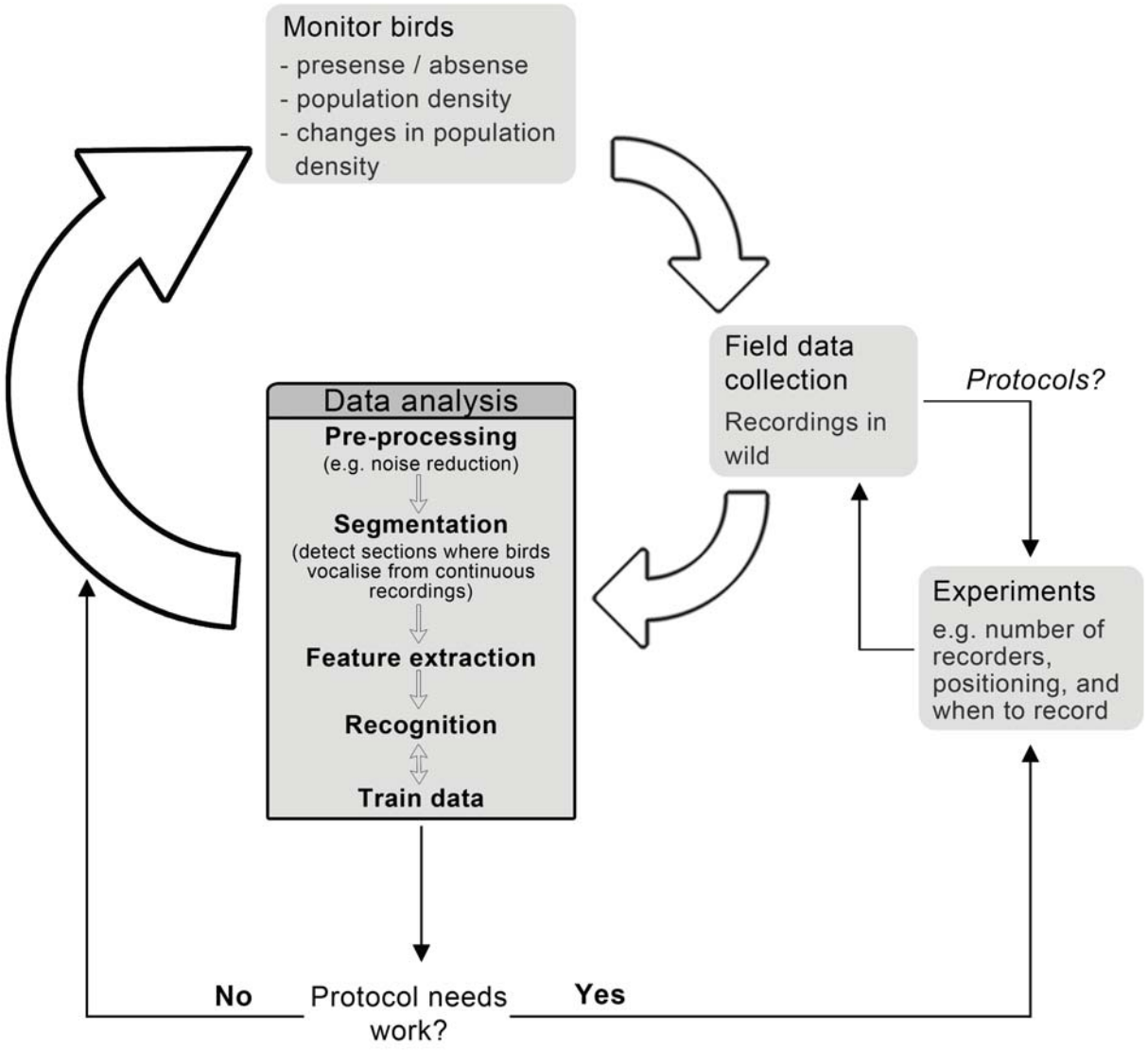


Figure 2.1: Representation of the full work process required for the use of acoustic recorders in wildlife management, including the development of protocols for the deployment of recorders.

set with clean references (less corrupted with noise) is helpful to avoid false positives (identification of any sound except the target bird sounds; e.g. Boucher (2014); Wildlife Acoustics, Inc (2011)). Usually, song examples from the same environment where the test recordings will be collected are preferred (Katz et al., 2016), as it hoped that these will have similar noise profiles and fewer variations in calls. Recording clean close-range calls is possible with handheld (manual) recorders because the recordist can get close to the individual birds and avoid noise by careful screening (Ruse et al., 2016). Although both manual and automatic recorders can be used to develop the birdsong database, it is often better to use birdsong made with programmable recorders if the test data are also collected with those recorders. The technology is slightly different between them: manual recorders are usually directional, meaning that they are more sensitive to the sound in one particular direction, which is ideal when recording individual birds. In contrast, the automated recorders use omni-directional microphones because they aim to record birds from anywhere around the recorder (Brandes, 2008). Unattended recorders produce higher quality recordings when the bird is close to the recorder and there is little noise in the environment.

The ground truth of the recordings is essential to be able to measure the performance of the recogniser (see Section *Performance Measures*). The labels and annotations need to be generated manually by experts by careful listening to the recordings and/or visual inspection of the spectrograms, a time-consuming but crucial task that is currently the main way that the processing of recordings is done. After the recogniser is trained and evaluated on known field recordings, it is ready to process unknown recordings, for example to identify the presence of a species.

2.2.2 Visual Representation of Birdsong

Sounds are generally visualised in two ways: the oscillogram or waveform (a plot of amplitude against time), and the spectrogram (a plot of frequency against time). The amplitude-time representation is the basic form of acoustic data. It can be turned into another useful representation by computing the power spectrum, which transforms the time domain signal into a frequency domain signal using short-time Fourier analysis. Two assumptions behind Fourier analysis are signal stationarity and periodicity. However, audio signals such as bird calls do not satisfy these assumptions, therefore, spectral analysis needs to be performed over very small periods of time (e.g. <100 ms). The spectrogram is created using these short-time Fourier transforms, generally computed from overlapping windows (for example, taking windows of 100 ms, spaced every 50 ms). First, the power spectrum is calculated, then rotated 90° and the amplitude is replaced by a greyscale (or colour) bar, with darker colours representing higher energy. The complete spectrogram is generated by stacking these bars along the time axis.

Different birdsongs generate different shapes and patterns in the spectrogram (Box 2.2; Fig. 2.2) and therefore spectrograms are commonly used by researchers and managers to visually scan recorded birdsong and to locate sounds of interest: spectrogram scanning and identifying target bird sounds by sight and the matching of spectrogram shapes is significantly faster than listening to the recordings.

Bird vocalisations are sometimes split into bird calls (fairly simple sounds produced by both sexes) and birdsong (long and complex sounds mainly produced by male songbirds (order *Passeriformes*)). The basic unit of all these vocalisations is the element. One or more elements together make a syllable and a series of syllables produce a phrase, these units are shown in Fig. 2.2.

Box 2.2: Some examples of shapes on spectrograms from bird calls

Lines: When bird calls appear as horizontal lines in the spectrogram we recognise them as 'whistles' (sounds with a continuous tone/frequency). Vertical lines represent 'clicks'. Angled lines portray 'slurs', which are sounds that cover many frequency modulated tones, from 'whips' to slow 'chirps'.

Warbles: These are a different kind of line segment. Within a warble a tone can be modulated in one direction and then back again.

Blocks: These are concentrations of acoustic energy that create some shape, such as a rectangle or triangle.

Stacked harmonics: These appear as a vertical stack of lines or warbles, often equally spaced. Usually the bottom of a stack represents the fundamental frequency, which generally has higher energy than the harmonics that appear above it.

Oscillations: These consist of a repeated acoustic component, typically a repeated 'click' or stacked harmonic.

2.2.3 Noise Reduction

Field recordings include any sounds that are present in the geographical area where the recorder is mounted, including birdsong of interest and many other biophony (sound from other bird species and animals), geophony (wind, rain etc.) and anthrophony (man-made sound, such as aeroplanes, and wind turbines). Therefore, finding the target sounds in a recording inbetween all the other noises is a challenge. Removing the 'noise' prior to birdsong detection and recognition is also constrained by the temporal and frequency overlap of the noise and the birdsong. Birdsongs overlap with each other as well as other noises; while the *niche hypothesis* states that birds spontaneously try

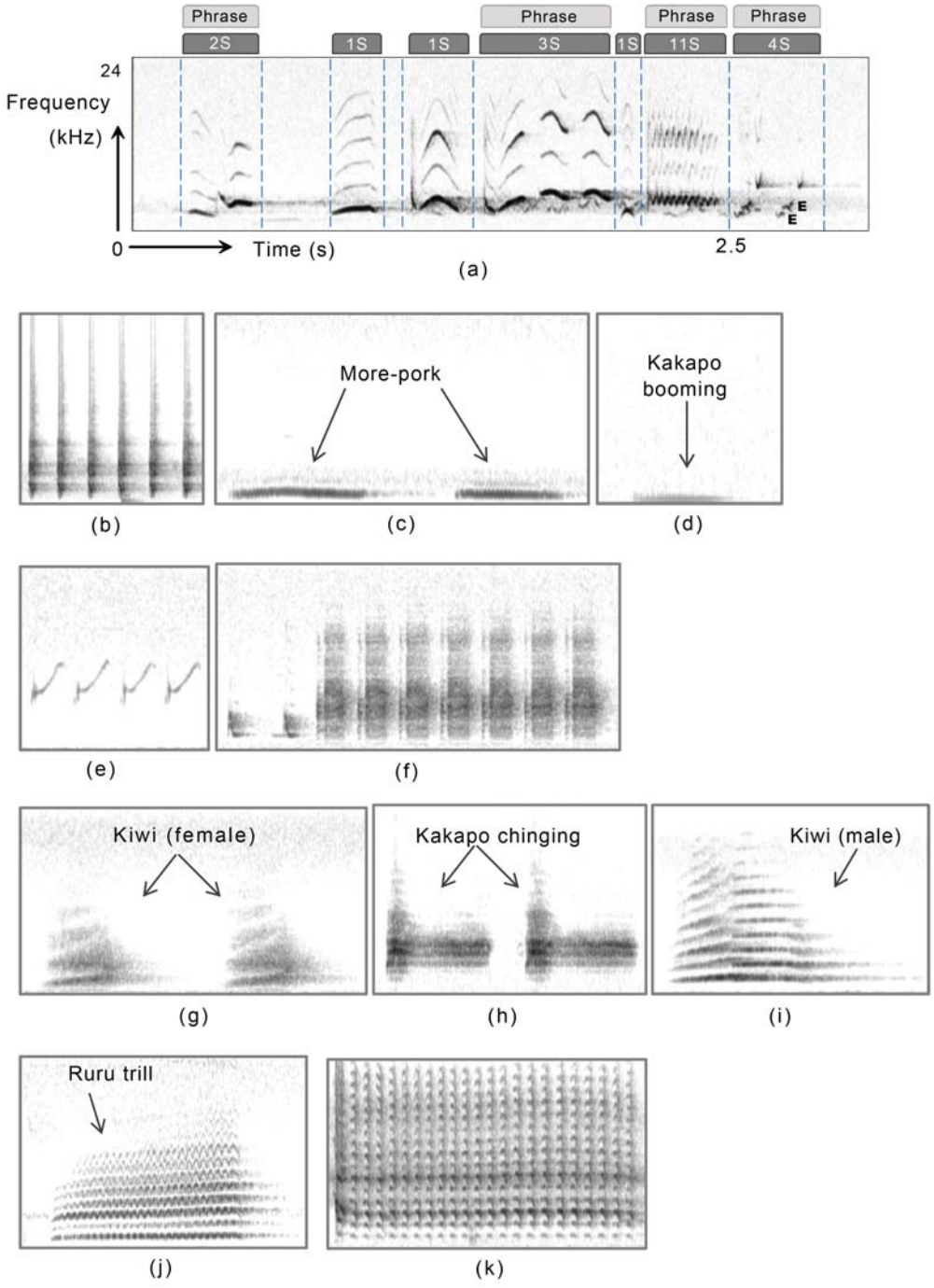


Figure 2.2: Spectrogram representations of various bird species showing some of the typical appearances of sounds. (a) A fox sparrow (*Passerella iliaca*) song illustrating its syllables, phrases, and elements (S=syllable and E=element). (b)-(e) show representations of lines: (b) tui (*Prosthemadera novaeseelandiae*); (c) the *more-pork* sound of morepork (*Ninox novaeseelandiae*); (d)kakapo[1](*Strigops habroptilus*) *booming*; (e) brewer's sparrow (*Spizella breweri*). (f)-(h) demonstrate blocks: (f) (long billed) marsh wren (*Cistothorus palustris*); (g) female North Island brown kiwi (*Apteryx mantelli*) call; (h) kakapo[1] *chinging*. (i)-(j) show stacked harmonics: (i) male North Island brown kiwi whistles; (j) morepork *trill*. (k) oscillations: North Island saddleback (*Philesturnus rufusater*).

[1] kakapo recordings from the Kakapo Recovery Team and Bruce C. Robertson

to avoid temporal and spectral overlap with other bird communities (Pijanowski et al., 2011b), examples that break this rule are not uncommon (Potamitis et al., 2014).

All these noises tend to hide or alter the birdsong; noise is superimposed onto the birdsong, degrading the signal quality and making the sound of the call fainter, leading to poor results by recognisers. Low recognition accuracy is often attributed to noise (Fox et al., 2006; Aide et al., 2013; Baker and Logue, 2003) because it affects the whole process unless removed initially. Relatively low frequency from abiotic sounds such as wind, aeroplanes, running water, and vehicles passing is common, and can either mask species with low frequency songs or generate false positives (Potamitis, 2014). Potamitis et al. (2014) stated that the overlap of unwanted sounds with the song of their two target species was the biggest obstacle in their automated processing of natural field recordings, while Wolf (2009) found that applying existing bird sound recognition methods to their raw recordings collected from forests was unsuccessful due to noise. Schrama et al. (2007) claimed that the overall recognition accuracy in automated recognition can only be achieved through an advanced denoising approach.

The most common approach for denoising found in the literature is to follow standard signal processing techniques and perform noise profiling, followed by filtering. However, noise profiling and noise filtering have their own limitations (Table 2.1). There are situations where high-pass, low-pass, or band-pass filters cannot be applied successfully. For example, if a particular species produces only low frequency songs, then a low-pass filter with a suitable cut-off frequency removes frequency components beyond the birds' frequency range. However, if there are multiple bird species or a species that produce vocalisations that lie in different frequency ranges, applying a filter without eliminating some bird vocalisations is impossible. The Wiener filter, an approach that is useful in signal enhancement to remove linear distortions (Vaseghi, 2008), is not useful for birdsong because it assumes that the signal and noise are stationary and that spectral information is available. While this can be partially overcome by using the spectrogram window method to split the signal into a series of small time-frames and compute the filter coefficients in each frame, this adaptive Wiener filter (Chen et al., 2006) suffers from signal distortion. So while it has been used for noise reduction in speech signals, it is not useful for birdsong recordings without *a priori* knowledge of the characteristics of the signal and noise, and this knowledge is rarely available with field recordings.

In recent work Priyadarshani et al. (Chapter 4 of this thesis; 2016) have shown that wavelets can be effectively used to remove noise from field recordings collected with automatic sound recorders. In a field recording, while the birdsong is transient, a

Table 2.1: Noise reduction methods in descending order of effectiveness in regards to automatic analysis of recordings. Effectiveness was decided based on the type of recordings used and the performance of the methods. A=automated; M=manual; L=long (≥5 min); S=short; SNR=signal-to-noise ratio; SnNR =modified SNR; AUC=area under the ROC curve; dB=decibles; Y=yes; N=no; N/G=not given.

Study	Method	Type of Recordings				Performance		
		A/M	No of species	L/S	Total length	Direct evaluation	Indirect evaluation	Species-specific?
Priyadarshani et al. (2016)	Wavelet denoising	A M	11	S	700 syllables and 55 complete songs	SnNR improved from 0.5 to 5.3 dB	Illustrated visually using examples	*N*
Ren et al. (2008)	Perceptually scaled wavelet packet decomposition	M	3	S	30 calls – manually added noise	N/G	Similar to uniform band wavelet packet transform	*Y*
Fox et al. (2008)	Band-pass filtering	M	3	S	N/G	N/G	87% individual identification accuracy	*Y*
Selin et al. (2007)	Modified filter bank (eight-band)	M	8	S	3,132 sound files	N/G	Main focus was species recognition (Table 2.3)	*Y*
Chu and Alwan (2009)	Correlation-maximization and band-pass filter	M	5	S	42 min	N/G	Classification error rate improved from 5.4 to 4.6	*Y*
Lasseck (2013)	Median clipping plus image processing on spectrogram	M	87	S	∼ 2 h	N/G	91.7% AUC – not reversible	*N*
Wolf (2009)	Wiener filter plus image smoothing (Gaussian blur, median)	A	1	S	single syllable	N/G	Illustrated visually using one spectrogram	*N*
Briggs et al. (2012)	Whitening filter (using a noise profile)	A	13	S	91 min	N/G	AUC of 0.978 (Table 2.3)	*N*
Baker and Logue (2007)	Noise profiling (subtract a sample noise)	M	1	S	1 sec×2 e.g.×4 manually added noises	SNR improved from 16 to 60 dB	Only useful when the noise is constant	*N*

considerable amount of background noise, particularly the geophony, is nearly stationary. Wavelet denoising eliminates this quasi-stationary noise no matter whether it is wideband or narrowband, providing that it is approximately Gaussian.

An alternative to audio analysis is treating the spectrograms as images and applying image analysis methods to clean them (Potamitis, 2014). While this method is useful to detect regions of interest (as will be discussed later), it is not reversible because the method does not permit the construction of a denoised sound file from the cleaned spectrogram. This means that it is of limited utility for recognition in general, as it is common to extract features from the sound file, as well as the spectrogram.

The result of successful noise reduction is a cleaner recording (although possibly with some artefacts) that is ready to be used as input into segmentation and recognition algorithms. We will consider those methods after we become familiar with the accuracy measures that are used to evaluate them.

2.2.4 Performance Measures

There are four possible outcomes when a classifier system makes an prediction for a binary output (such as whether or not there is a birdcall in a given segment of recording): a true positive (TP) is when the classifier correctly says that there is a birdcall, a true negative (TN) is when it correctly says that there is not, while false positives (FP) and false negatives (FN) are where the classifier incorrectly suggests that there are and are not (respectively) birdcalls. The number of examples of each of the four cases will add up to the number of segments considered. Note that this assumes that the segmentation algorithm took in short time segments and processed each individually. It may be that, instead, it identified the start and end times of each birdcall. In this case, the same four outcomes can be considered, but the base numbers will be different.

The outcomes can be combined into a few different useful measures:

Recall (also known as sensitivity or the true positive rate) $= \frac{TP}{(TP+FN)}$

Precision $= \frac{TP}{(TP+FP)}$

Specificity $= \frac{TN}{(TN+FP)}$

Accuracy $= \frac{TP+TN}{(TP+FN+FP+TN)}$

$\boldsymbol{F_\beta} = \frac{(1+\beta^2) \times (Recall \times Precision)}{(Recall + \beta^2 \times Precision)}$, where β is a weighting factor (often $\beta = 1$ or 2)

For example, the strength of a call detection algorithm is commonly determined by its recall rate and precision. The recall rate explains how well the segmentation algorithm detects songs, or the percentage of songs retrieved from the total number of songs in the recording. Precision refers to how reliable the detection algorithm is (the percentage of true positives from the positively classified songs). Maintaining a high recall rate while achieving high precision is challenging: recall and precision are often inversely related to each other: it is possible to gain recall at the cost of losing precision and vice versa. Specificity measures how good a given model is at avoiding false alarms, while accuracy refers to the proportion of the total number of predictions that were correct. Sometimes a combined version of recall and precision, named the F_β score, is used to express the balance between recall and precision.

One way to present some of these measures is to use the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR=recall) against the False Positive Rate (FPR=1-specificity) as they change when some parameter(s) of the method are varied. For example, an ROC curve could be generated as a threshold at which sounds are considered as signal rather than noise is adjusted (normally with the effect of increasing TP at the cost of an increased FP or vice versa) producing different (TPR, FPR) pairs to be plotted. The Area Under the (ROC) Curve (AUC) is also used as a measure of accuracy (for further information, see Marsland (2014)).

The same measures can be used to evaluate classification algorithms, except that there are usually more than two species of bird to be identified. The definition of a false positive is then modified to be an incorrect choice of class. While the majority of researchers have used the aforementioned metrics some have used different measures. It is not always clear from papers precisely what measures were used, hence Tables 2.1–2.3 report the performance of each method only to the best of our knowledge and understanding of the relevant materials provided.

2.2.5 Call Detection and Segmentation

In general, automatic recorders turn on and off at set times, and record everything between those times. In order to recognise the birdsong in the recording, potential sections that could contain calls need to be isolated first. This can also be useful to enable human analysts to concentrate only on what is important in a long recording. This is even more important when the recordings that are collected with scheduled recorders contain very little birdsong, but lots of noise: useful data are then a very small proportion of the total recording (Andreassen et al., 2014). Selin et al. (2007) concluded that call detection and segmentation is the most complicated and difficult part of the whole automation process; they also highlighted the need for noise reduction.

Following call detection, it may also be desirable to divide a song or a series of

calls into syllables, and this is also not straightforward (Tchernichovski et al., 2000), especially when the syllables are not followed by a silent interval and not separated clearly. Merging the syllables that are very close to each other is common practice in syllable detection (Fagerlund, 2004). Further, isolation of acoustic units also poses a great challenge due to background noise. Therefore, the majority of the published work has largely avoided automatic segmentation and instead used manually segmented data (Somervuo et al., 2006; Anderson et al., 1996; Franzen and Gu, 2003; Fox et al., 2008; Chen and Maher, 2006) to test their recognition methods. Even when automatic segmentation is used, it is sometimes followed by manually eliminating noisy segments (Selin et al., 2007).

Both the waveform and the spectrogram can be used to isolate bird vocalisations (Table 2.2), based on the assumption that the sections where the birds sing carry more energy than the other parts of the recording (Jinnai et al., 2012; Harma and Somervuo, 2004; Somervuo et al., 2006; Towsey et al., 2012; Juang and Chen, 2007; Murcia and Paniagua, 2013). This assumption is valid for recordings that are not corrupted by too much noise, but this is not always the case for automatic recordings. Noise causes the bird sounds to be quieter and faded (Briggs et al., 2012), adding to issues with bird proximity to the recorder. Common energy-based song detection coupled with thresholding fails to detect faint songs, but also detect periods of noise that exceed the chosen threshold.

The most common frequency-based method is to treat the spectrogram as an image and use median clipping, whereby points are identified as birdsong if they are more than some pre-defined multiple of the median of the relevant column and row of the spectrogram (Fodor, 2013; Lasseck, 2014; Potamitis, 2014). Image processing techniques such as basic shape morphology methods can be used to improve this process (Potamitis, 2014).

Very few studies have evaluated detection methods on natural unattended field recordings (Jančovič and Köküer, 2015; Bardeli et al., 2010; Frommolt and Tauchert, 2014). Jančovič and Köküer (2015) applied a sinusoidal detection method assuming that the number of species in a given recording is known. The recall decreased significantly (from 80% to 71%) when the number of species increased from 2 to 3. Bardeli et al. (2010) made some valuable contributions to the field by proposing two algorithms specialised to detect two bird species. The methods are based on the temporal patterns in the frequency bands of the target species and noise estimation from each band followed by spectral subtraction to avoid noise, and by evaluating the methods on a large set of automated recordings (26 hours). They reported 94% recall and 66% accuracy on the Eurasian bittern (*Botaurus stellaris*) and 92% overall detection rate on Savi's warbler (*Locustella luscinioides*). The problem with these species-specific methods is that

they need to be completely redesigned in order to detect other bird species. Frommolt and Tauchert (2014) used template matching in *Avisoft-SASLab Pro* (Specht, 1993) to detect Eurasian bitterns in controlled field recordings (data collected under calm conditions) and reported 85% recall with no false positives.

2.2.6 Feature Choice and Extraction

Turning the denoised and segmented birdcall into something suitable for input into a classification algorithm requires that features of importance are extracted from the call. Possible features can be as simple as the sequence of amplitudes present in the call (or the raw spectrogram values), but generally more success is found with more descriptive features; the overall recognition performance of any pattern recognition task depends heavily on the suitability of the features considered. There have been a huge number of features considered in the literature, often based on those considered in other areas, such as speech recognition and music processing. In general, if it is unclear which features will be helpful, the tendency is to add more in. Unfortunately, the more features that are included, the more training data are required for learning, something that is known as *the curse of dimensionality* (Marsland, 2014).

There are several ways to categorise the features that are derived from sounds. In general, only local information (based on individual short time windows) is useful for birdsong, particularly since the calls are segmented from the recording. Features can be based on the amplitude plot (such as the bandwidth, number of zero crossings), or on the energy of the signal within the window, or on the short-time Fourier transform data (such as its statistical moments, fundamental frequency, or spectral variations). Many of these features are related to one another despite being based on different representations. For example, pitch and loudness of a bird call are related to the frequency and energy (cumulative amplitude effect over time) respectively.

One set of interesting features use the short-time Fourier transformed data and transform it to more closely match how humans process sound. These so-called perceptual features can be based on either the Bark scale or the Mel scale. One set of features that are particularly common in the literature are Linear Predictive Coding (LPC) and its extension, the Linear Prediction Cepstrum Coefficients (LPCC), that were initially used for encoding human speech (Zbancioc and Costin, 2003), but also seem to be useful to represent birdsong (Table 2.3). Mel Frequency Cepstral Coefficients (MFCC) use the Mel scale filter bank (Graciarena et al., 2010), which consists of logarithmically spaced Mel scale filters as well as linearly spaced filters (as a collection of overlapping band-pass filters), linear frequency spacing below 1 kHz and log frequency spacing above 1 kHz. MFCC are derived by applying the Discrete Cosine Transform (DCT) to the logarithm of the energies, producing a relatively low-dimensional representation.

Table 2.2: Call detection and segmentation methods in descending order of their effectiveness to automatic detection of putative calls. Effectiveness was decided based on the performance and the type of recordings used. A=automated; M=manual; L=long (≥5 min); S=short; N/G=not given; AUC=area under the ROC curve; SNR=signal-to-noise ratio.

Study	Method	Recordings				Performance	Comment
		A/M	No of species	L/S	Total length		
Bardeli et al. (2010)	2 species-specific algorithms – based on the temporal patterns in the frequency bands – noise estimation and spectral subtraction	A	2	L	26 h	94% recall, 66% accuracy on Eurasian bittern – 92% detection rate on Savis warbler	Demonstrated on two very different species
Frommolt and Tauchert (2014)	Template matching by spectrogram correlation using *Avisoft-SASLab Pro* (Specht 1993)	A	1	L	5 h	84.9% recall, 100% precision (assuming the software output raw detections)	Species-specific – needs testing with noisy field recordings
de Oliveira et al. (2015)	Morphological opening (erosion and dilation) – HMM for recognition	A	1	L	∼ 3 h	56% recall	Required large amounts of training data (> 27 h)
Lasseck (2013)	Median clipping followed by standard image processing techniques	M	87	S	∼ 2 h	91.7% AUC (NIPS4B 2013 competition)	Needs evaluation against long continuous (long) recordings
Jančovič and Köküer (2015)	Sinusoidal detection method	A	30	S	33 h	80% recall with 2 species and 71% recall with 3 species	Assumes the number of species in recordings is known – does not appear to scale well
Potamitis et al. (2014)	Hilbert follower – down sampling and band pass filtering	A M	2	S	>42 h	RMBL-Robin: 91% recall and 71% precision – Vouliagmeni (kingfisher): 85% recall and 85% precision	Authors focused on reducing the search space by discarding recordings devoid of target species calls

Ranjard et al. (2015)	Manually defined 5 HMMs to model target species, other species, humans, recorder noise, and silence – MFCC and PLPC features – NN classifier	M	1	L	24 min (test) – ∼28 h (full dataset)	100% recall and 78% precision on test file	Not scalable
Jinnai et al. (2012)	Time domain energy curve thresholding	A	1	L	N/G	N/G	Their software has reported more than 95% accuracy on dawn chorus (Boucher et al., 2012)
Schrama et al. (2007)	Frequency band thresholding to extract flight calls	A	N/G	N/G	12 h	27% recall	Emphasised the need of prior noise removal
Neal et al. (2011)	Random forest classifier	A	N/G	S	625 sound files	93.6% recall – 8.6% false positive rate	Original continuous recordings were sampled (15 s) to generate dataset
Zhang and Li (2015)	Adaptive energy detection – SVM classifier – Mel-scale and wavelet features	M	N/G	S	30 bird sounds	85% classification accuracy	Used clean recordings manually polluted with noise
Harma and Somervuo (2004)	Time domain energy curve thresholding to detect syllables	M	150	S	2,000 record-ings	Sinusoid model matched with spectral structure of 93% syllables	
Somervuo et al. (2006)	Time domain energy thresholding to detect syllables	M	14	S	792 birdsong	85% recall – 93% precision	High variability in accuracy between species
wa Maina (2015)	Speaker diarisation techniques	M	19	S	179 record-ings	42.5% accuracy of estimating the number of species in recordings	Assumed one cluster represents all the non-bird sounds
Harma (2003)	Sinusoidal modelling to decompose birdsong into syllables and identify them	M	14	S	N/G	38.7% overall recognition rate	Needs estimate of number of syllables
Lee et al. (2006)	Modified Harma (2003)	M	420	S	420 sound files	87% recall	Classification errors are to poor segmentation and/or background noise
Juang and Chen (2007)	Time domain energy	M	10	N/G	N/G	94.67% recognition rate	Method only useful when the SNR is high

As MFCC has been useful for human speech recognition (Makhoul and Schwartz, 1995; Muda et al., 2010; Priyadarshani et al., 2012), researchers have used MFCC to represent features in animal vocalisations (Lee et al., 2006; Briggs et al., 2009; Stattner et al., 2013; Kogan and Margoliash, 1998; Fox et al., 2006; Clemins and Johnson, 2003). Mostly, MFCC are used with their first and second order derivatives in order to capture dynamic features of the vocal tract. While LPCC is a low cost approach, MFCC has proven to be more accurate for classifying animal noises (Lévy et al., 2003). However, there are contrasting views regarding the sensitivity of MFCC to noise: Singh et al. (2012) report that MFCC are less susceptible to additive noise, while Wu and Cao (2005) find the opposite.

Given this plethora of possible features and the fact that using more features requires more training data, and can lead to less accurate results, it is necessary to find subsets of the features that are most useful. Accordingly, the goal of feature selection is to identify redundant or unnecessary features that can removed in order to reduce the input dimensionality while retaining most of the information, thus enabling more accurate classification. Principal Component Analysis (PCA) is a useful tool in this regard, and can effectively reduce the data dimensionality. For example, Somervuo and Härmä (2003) significantly reduced the dimension of their birdsong feature vectors from 1,000 to 7, with 99% of the variance being explained by those seven components. Another approach, based on data mining of the spectral features was used by Vilches et al. (2006, 2007). An alternative approach is to perform a pre-classification by identifying windows that are similar, so that fewer of them are used. Vector Quantization (VQ) can be used for this by clustering together similar windows.

While Fourier analysis is the foundation of most of the previously mentioned frequency-based features, the short-time Fourier transform suffers from a lack of time-frequency resolution. Wavelet analysis is an alternative (Priyadarshani et al., 2016). In both the Fourier transform and the wavelet transform, a given signal is converted into frequency domain, but the difference is in the basis functions: the Fourier transform is based on sinusoids, while wavelets are based on self-similar basis functions called mother wavelets. In contrast to sinusoids, wavelet functions are localised in space and are scale-invariant. Therefore, the trade-off in time-frequency resolution can be avoided by using large windows for low frequencies and small windows for high frequencies simultaneously. Bastas et al. (2012) observed that Discrete Wavelet Transformation (DWT)-based features outperformed MFCC. Despite the opportunities that wavelets provide, relatively few publications have so far used them for birdsong analysis (Selin et al., 2007; Turunen et al., 2006; Chou and Liu, 2009; Zhang and Li, 2015).

Regardless of what sounds are being studied, audio feature extraction is a common

topic in audio signal processing. As such, a large number of toolboxes are readily available to easily extract widely used features, particularly for speech and music. Moffat et al. (2015) provides an evaluation of the major feature extraction tools. Despite all of this research there is as yet no clear evidence for and against different representations for birdsong with many different approaches being used (see Table 2.3).

As bird calls are temporal, there are two further processing challenges that have to dealt with for feature selection: window size selection and temporal alignment. The first of these requirements is due to the fact that many of the features that have been found useful for birdsong recognition are based on short time windows, and establishing a suitable window size is important, in both time and frequency range. For field recordings, where there can be birds with a wide range of different pitches and lengths of calls, both of these choices can be tricky: longer time windows have lower time resolution but higher frequency resolution, and vice versa. Graciarena et al. (2010) investigated the generalising capacity of MFCC over 92 different bird species in conjunction with a Gaussian Mixture Model (GMM) and found that the optimum frame length is species specific. However, they found that frequency range optimisation is possible even though the species do not share the same frequency band (100–13,000 Hz was selected experimentally). In their experiment, the best number of filters in the Mel filter bank was 41, which is high compared to speech (generally 13 filters). Chu and Alwan (2012) proposed an algorithm to optimise the filter bank parameters by using an Expectation Maximisation (EM) algorithm. Using a 42 minute recording as test data with an 85 minute recording as training data, Chu and Alwan (2012) improved the identification error rate from 8.7% to 6.2% on the calls of five antbird species.

The second challenge is to align the calls within the window, a process known as temporal alignment. The most common method for performing temporal alignment is Dynamic Time Warping (DTW), initially proposed by Vintsyuk (1971) for automatic word recognition. It has been successfully used by many researchers to match birdsong (Kogan and Margoliash, 1998; Somervuo et al., 2006; Anderson et al., 1996). DTW successfully copes with the different birdsong speeds and lengths by stretching and squashing sections of birdsong so as to find the best matching alignment (Coleman, 2005).

2.2.7 Recognition and Classification

Reproducing human-level processing of sight and sound is one of the holy grails of machine learning. However, despite recent advances, we are still a long way from this. Machine learning methods generally take vectors of equal length and compute representations of them in order to cluster similar inputs together. The features that comprise the vector are typically based on some subset of the methods described in

the previous section. The challenge is to find a representation of the feature vectors that makes the examples of one particular type of call from one species, in all their variation, similar to each other, but dissimilar to other calls, and all calls of any other species.

Once a feature representation has been chosen, feature vectors extracted from the sound file can be fed into a standard machine learning algorithm, which will cluster those that are similar, either in an unsupervised fashion (i.e., without human labels for the calls) or using labels provided by human experts beforehand (supervised learning). Alternatively, exemplars of each call can be treated as templates of a particular type of call, and the distances between vectors representing each call and a new call can be computed, with the closest exemplar being declared a match to the new call.

There are a plethora of machine learning algorithms, and we will only mention those that have been used in the literature for birdsong recognition (see Table 2.3). However, we will first summarise some possible distance metrics between vectors, as this will be important for comparing some of the methods.

Distance Metrics

One approach is to treat the vector as describing a position in a high dimensional space, for example a feature vector with 16 elements means that is a 16 dimensional space; often the feature vectors are rather longer than that. In that case, the common distance metrics between points can be used; the Euclidean distance, which is a particular example of a Minkowski distance (Marsland, 2014) is one example of this type of metric. However, for high dimensional data these methods do not deal well with noise. The alternative is to treat the feature vectors as samples from some unknown probability distribution and seek a match there, for example by using the Kullback-Liebler (KL) divergence. This is a divergence, not a distance, since it is not symmetric: in general $\mathrm{KL}(A \parallel B) \neq \mathrm{KL}(B \parallel A)$.

Alternative distance measures can be derived for particular applications, and one that is used for birdsong recognition is the geometric distance (Jinnai et al., 2012), which is the basis for SoundID (see the section *Current Birdsong Recognition Related Software*). This metric aims to be robust to noise by comparing the vectors with a Gaussian distribution and measuring the kurtosis. A more explicit use of the Gaussian distribution is the statistical learning method known as a Gaussian Mixture Model (GMM). It is assumed that observations (such as recorded bird song) come from a weighted combination of inputs that can be described by Gaussian random variables. As some of the input can be noise as well as types of bird call, this method can deal well, at least in theory, with multiple noise sources. One challenge is knowing how many sources there are, and training them appropriately.

As with the feature selection part of birdsong recognition, many machine learning-based approaches have taken their lead from automated speech recognition, even though the two contexts have as many differences as commonalities (Doupe and Kuhl, 1999); examples include (Skowronski and Harris, 2006; Zhang and Li, 2015; Somervuo et al., 2006). In particular, the sounds that birds make are far more variable than those of humans, and the environmental conditions of the recordings are very different.

The most commonly-used form of machine learning for birdsong recognition is the Neural Network (NN). However, there are two particular problems with this method: it acts as a black box and usually scales badly, so that its ability to discriminate between species degrades as a larger variety of calls is introduced (Cai et al., 2007). Hence, while neural network methods work fairly well for limited number of species, this falls off as the number of species increases; for example, Lopes et al. (2011a) started to classify 73 species, but concluded that the performance varied considerably between the species and suggested that a maximum of 12 species could be used within the F_1 measure margin of 80%.

Neural networks that perform unsupervised learning, meaning that they cluster similar calls together rather than using human labels to recognise calls from the same species, are also commonly used, particularly the Self-Organising Map (SOM). Although Stowell and Plumbley (2014) suggest that unsupervised learning is key to successful recognition, Selin et al. (2007) observed higher recognition accuracy with a supervised method (the Multi-Layer Perceptron (MLP)), obtaining 96% test accuracy against 78% using SOM when recognising eight bird species.

The Support Vector Machine, a machine learning approach that maps data into a higher dimensional space where it can be linearly separated, has been used in many of the studies in Table 2.3. While it often produces very impressive results, as the amount of data increase, so the computational costs increase, and this makes it unsuitable for processing large numbers of calls.

There has also been interest in using decision trees, which assemble a tree that can be interpreted by humans: at each node of the tree a split is performed using just one of the features from the input vector, and a sequence of these splits enables a decision as to species to be made at the leaves of the tree. A collection of decision trees that are created using random partitions of the data and that independently produce outputs that are combined by majority voting is known as a random forest. They are relatively easy to train and use, and have shown positive results for birdsong recognition in a species of kiwi (Digby et al., 2013), and also larger number of bird species (Lasseck, 2015a; Fodor, 2013; Potamitis, 2014).

Table 2.3: Feature extraction and recognition in the descending order of their effectiveness to automatic recognition of bird sounds. Note that the first few rows are allocated to the series of recent birdsong recognition competitions; there is no evidence that these would actually be useful for recognition from automatic recorders. A=automated; M=manual; L=long (≥5 min); and S=short; AUC=area under the ROC curve; dB=decibels; GB=gigabytes; CD=compact disk; N/G=not given.

Study	Features used	Recognition method	Type of recordings				Performance
			A/M	No of species	L/S	Total length	
Murcia and Paniagua (2013)	MFCC – feature reduction by linear discriminant analysis	NN	M	35	S	3.75 h testing – 18 min training	Winning solution of ICML 2013 technical challenge – 0.74 AUC
Dufour et al. (2013)	MFCC	SVM	M	35	S	3.75 h testing – 18 min training	0.64 AUC in ICML 2013 technical challenge
Fodor (2013)	Spectral features were manually selected (30-70) for each species	Random forests	M	19	S	54 m testing – 54m training	Winning solution of 2013 MLSP – 0.956 AUC (23 random forests) – 0.955 AUC (1 random forest)
Lasseck (2013)	File statistics – segment statistics – segment probabilities	Decision trees	M	87	S	∼ 2 h – 687 train and 1,000 test recordings	Winning solution of NIPS4B 2013 – 0.92 AUC – segment probabilities were more useful
Potamitis (2014)	Descriptive, morphological features, and spectrographic cross-correlation	Random forests	M	87	S	∼ 2 h – 687 train and 1,000 test recordings	0.91 AUC in NIPS4B 2013 competition
Lasseck (2014)	6,669 low level descriptors including MFCC – reduced to 1,277 after feature elimination	Decision trees	M	501	S	9,688 train and 4,339 test recordings	Winning solution in LifeCLEF 2014 Bird Identification Task – 0.92 AUC – 51% recall

Lasseck (2015a)	8,541 low level descriptors including MFCC – best results were gained with 500 features	Decision trees – bootstrap aggregating to combine multiple classifiers	M	999	S	33,862 recordings	Wining solution in LifeCLEF 2015 Bird Identification Task – 0.97 AUC – 45% recall
Ganchev et al. (2015)	MFCC (without Mel-filter bank)	A statistical log-likelihood ratio estimator based on the GMM universal background model and post-processing	A	1	L	2 h 20 min test dataset – >27 h train dataset	Recall: 97.7% (0,-20 dB), 83.8% (0,-30 dB), 43.3% (0,-40 dB), and 31.2% (0,-50 dB) – 100% precision – used to detect Southern Lapwing (*Vanellus chilensis lampronotus*) from one month (24/7) recordings
de Oliveira et al. (2015)	Bird sound detection using morphological filtering of the spectrogram	HMM	A	1	L	2 h 20 min test dataset – >27 h train dataset	Recall (56%) was better than GMM based segmentation (55%; Sahidullah and Saha, 2012) and syllable based method (44%; Harma, 2003)
Potamitis et al. (2014)	Perceptual LPCC, a slightly different version of MFCC	HMM	A M	2	S	>42 h	RMBL-Robin :77% recall and 85% precision; Vouliagmeni: 85% recall and 85% precision – separate models for target and noise
Chu and Blumstein (2011)	–	HMM	M	1	S	∼78 min (RMBL-Robin)	76% recall and 75% precision – separate HMMs to model target species and background sounds
Acevedo et al. (2009)	Time-frequency features – linear discriminant analysis	Decision trees – SVM	A	12	S (1 min)	>35 h	95% recall and 99% precision – SVM was better than decision trees
Bastas et al. (2012)	DWT, MFCC, Spectrogram-based Image Frequency Statistics (SIFS), and Mixed MFCC and SIFS (MMS)	k nearest neighbours (kNN), MLP, HMM, Evolutionary Neural Network (ENN)	A	5	S	459 test and 128 train bird calls	94% accuracy – Song Scope resulted less accuracy (70%)
Ulloa et al. (2016)	A mean template derived from 10 standardised samples	Spectrogram cross-correlation	A	1	S (1 min)	5 h 36 min test dataset	35% recall and 100% precision – canopy level recorders (20 m) detected more calls (62%) than those placed lower (1.5 m, 38%)

Study	Features used	Recognition method	Type of recordings				Performance
			A/M	No of species	L/S	Total length	
Briggs et al. (2012)	The shape of the binary mask of the segments	Multi Instance Multi Label (MIML) SVM, MIML kNN, MIML Radial Basis Function (RBF)	A	13	S (10 sec)	548	MIML RBF reached the highest AUC (0.98)
Andreassen et al. (2014)	Species-specific features (duration, average power, spectral entropy)	SVM	A	1	S	59 GB	precision: 96% (when dry) and 70% (whan rain)
Digby et al. (2013)	5 species-specific features	Decision trees (C5.0)	A	1	L	52.2 h test – 66 h train	Automatic: 40% recall, 98% precision – manual scanning: 80% recall – field survey: 94% recall, 96% precision – 3 methods resulted similar annual change of calling activity
Jančovič and Köküer (2015)	Estimation of frequency tracks	HMM	A	30	S	33 h	Accuracy (78%) dropped to 69% when the number of species is unknown
Ventura et al. (2015)	MFCC – noise reduction and frame selection using morphological filter (considering spectrogram image)	HMM	M	40	S (avg. 32 s)	200 test and 400 train recordings	Precision (72%) was higher than MFCC after frame selection with GMM energy detector (70%; Sahidullah and Saha, 2012); syllable segmentation (65%; Harma, 2003); region of interest detector (48%; Potamitis, 2014; Briggs et al., 2012)

Vilches et al. (2007)	71 pulse features – data mining	HMM, Decision trees (VQ+ID3, J4.8), and Naive Bayes	M	3	S	204 recordings	Best accuracy (98%) from J4.8 (47 features) – incomplete vocalisations and background noise led to low accuracy of HMM (93%)
Selin et al. (2007)	4 features (maximum energy, position, spread, and width) derived from wavelet packet decomposition coefficients	SOM and MLP	M	8	S	2,278 train and 854 test sounds	High recognition accuracy with MLP (96%) compared to SOM (78%)
Fagerlund (2007)	MFCC and descriptive parameters	SVM	M	8	S	2,278 train and 854 test sounds	98% accuracy – similar performance as (Selin et al., 2007)
Papadopoulos et al. (2015)	Spectral mean, standard deviation, skewness, kurtosis, mode, and spectral flatness (Madhu, 2009)	GMM	M	15	S	N/G	Best results when using one feature (mode): >0.90 AUC for 11 species in 'park' category and 10 species in other cases – lowest AUC from same species (0.70, 0.64, 0.56 for three noise types)
Chen and Maher (2006)	Peak frequency track	Spectral distance (distance between the test frequency track and the reference) and threshold	M	12	S	12 test recordings	Accuracy (99%) decreased to 95% in presence of noise (manually adding white noise) – LPCC and DTW (90% to 71%) – MFCC and HMM (95% to 76%)
Wielgat et al. (2012)	MFCC	HMM	M	30	S	1,426 songs	92% recall – not tested on field recordings; separate model for each species
Brandes (2008)	Change in peak frequency and in bandwidth over time	HMM	M	9	S	5,871 test and 908 train calls	75-96% recall – cricket and frog species recognition was easier than birds
Taylor (1995)	Peak frequency track (single)	Decision trees (C4.5)	M	9	S	138 flight calls	78%, 4%, and 18% of calls identified correctly, incorrectly, and left unclassified respectively

Study	Features used	Recognition method	Type of recordings				Performance
			A/M	No of species	L/S	Total length	
Tan et al. (2012)	Normalised spectrogram (PCA for dimensionality reduction)	A sparse representation based classifier that represents test feature vector as a sparse linear combination of training data	M	1	S	1,022 syllables	90% accuracy – SVM and nearest subspace classifier resulted 88% accuracy
Dong et al. (2015)	Spectral ridge features	Euclidean distance	N/G	19	S	5 h	94% accuracy – when spectral ridges are strong method worked well, but not when the acoustic energy is temporally and spectrally diffused (e.g. parrot shriek)
Kasten et al. (2010)	105 features	Anomaly detection	N/G	10	S	3,673 segments	82% accuracy
Lopes et al. (2011a)	MFCC and descriptive parameters (MARSYAS framework; `http://marsyas.info`)	Naive Bayes algorithm, kNN with k=3, decision tree classifier (J4.8), MLP, SVM, Sequential Minimization Algorithm (SMA)	M	73	S	1,619 recordings	2 databases: one with complete audios and one with only pulses, the second was better – MLP and SMA performed better – maximum F_1 when 3, 5, 8, 12, and 20 species considered was 95%, 89%, 86%, 83%, and 78% respectively
Lopes et al. (2011b)	Compared MARSYAS feature set, Inset-Onset Interval Histogram Coefficients feature set, and Sound Ruler feature set	Naive Bayes algorithm, kNN with k = 3, decision tree classifier (J4.8), MLP, and SVM with SMA	M	3	S	101 songs	MARSYAS and Sound Ruler performed equally well – pulses database (99.7% F_1) were better than original audios (79.2%)

Ganchev et al. (2012)	Temporal and spectral features from openSMILE (Eyben et al., 2010)	kNN, Bayes network, MLP, C4.5 decision tree (J4.8), and SVM with SMA	M	N/G	S	150 test recordings – 12 min for training	Recognition accuracy: 86% (MLP), 81% (SVM) – classified each sound frame as bird sound or noise – not addressed species recognition
Ross (2006)	20 frequency, cepstral, and multi frame (global) features	MLP, SVM, Kernel Density Estimation of probability (KDE)	M	10	S	403 test and 193 train	79% precision – MLP had the highest accuracy (83%) – authors initially discarded noisy examples
Tyagi et al. (2006)	Spectral Ensemble Average Voice Print (SEAV) computed on FFT spectrum by frame wise averaging FFT coefficients	A comparison between SEAV + Euclidean distance, spectrogram + DTW, and MFCC+ GMM	N/G	15	N/G	63 recordings	Recognition performance (not defined): SEAV + Euclidean distance 87% – spectrogram + DTW 67% – MFCC+ GMM 100%
Graciarena et al. (2010)	MFCC	GMM	M	92	S	>=4 recordings per species	Minimum equal error rate (9 on train dataset and 10 on test dataset) was achieved with high number of filters (41) with 100-13,000 Hz
Heller and Pinezich (2008)	Frequency track extraction	Mahalanobis distance	M (high SNR)	4	S	N/G	79% recall
Kogan and Margoliash (1998)	MFCC	DTW and HMM	M (lab recordings)	2 (males)		993 songs	90% recall – DTW was better depending on the quality of recordings and complexity of songs
Trifa (2006)	MFCC	HMM using HTK	M	5	S	3,368 songs	Accuracy (99.5%) decreased to 90% after introducing low quality recordings
Fox et al. (2006)	MFCC	NN	M	3	S	24 recordings	89% precision – highly restricted experiments (1 recording per individual) – focused individual identification

Study	Features used	Recognition method	Type of recordings				Performance
			A/M	No of species	L/S	Total length	
Jančovič and Köküer (2011)	MFCC and tonal based features (frequency and magnitude of the prominent tonal component per frame)	GMM	M (CD quality)	99	N/G	N/G	83% recall – 1% precision – performance was reported at 10 dB SNR – recognition was in binary format (signal/noise) – tonal features were better than MFCC in the presence of manually added white noise
Mundry and Sommer (2004)	Estimation of fundamental frequency contour	Not addressed	N/G	2	N/G	N/G	No direct evaluation – fundamental contour differed in individuals – looked at the relationship between the structure of begging calls and nutrition need of chicks
Somervuo et al. (2006)	Combination of MFCC and descriptive parameters	DTW, GMM, and HMM	M	14	S	792 birdsong	Precision: 60% (song level); 40% (syllable level) – best results by using MFCC-based syllable trajectory models with DTW – recognition of GMM and HMM were almost similar
Lee et al. (2008)	Two dimensional MFCC	GMM	M (CD)	28	N/G	3,789 syllables	84% classification accuracy
Lee et al. (2013)	Angular Radial Transform (ART) on spectrogram	GMM	M (CD)	28	N/G	3,789	95% classification accuracy
Kwan et al. (2006)	MFCC	GMM	N/G	11	N/G	N/G	90% classification accuracy (SNR = 5 dB)

McIlraith and Card (1997)	Time-frequency information	NN	M (CD)	6	S	133 birdsong	>90% accuracy (not defined)
Cai et al. (2007)	MFCC	MLP	A	14	N/G	N/G	Recognition rate (not defined) stood just below 99% with 4 species, but decreased (87%) after introducing other 10 species
Somervuo and Härmä (2003)	Sinusoidal modelling	SOM and DTW	M	5	S	1,000 syllables	Divided the SOM map into five areas (equal to number of species)
Juang and Chen (2007)	LPC	An extension of NN	M (CD)	10	N/G	N/G	96% accuracy – manually segmented data – separate NN for each species
Lee et al. (2006)	LPCC and MFCC	Euclidean distance between test pattern and references	M (CD)	420	S	420 recordings	87% recall – MFCC was better than LPCC – wrong classifications were due to background noise and poor segmentation
Tachibana et al. (2014)	532 features including spectral and cepstral features	SVM	M	1	N/G	N/G	99% accuracy – the research contributes to neuroscience where syllable classification of one species is common

All of the methods that we have considered above use some form of feature vector constructed from a time window of the bird call. However, many birdsongs consist of multiple syllables put together. In order to recognise a sequence of syllables, another method that is commonly used in speech recognition has been applied: the Hidden Markov Model (HMM). This method creates a time-dependent probability distribution showing how likely certain syllables are to follow from others (Kwan et al., 2004; Kogan and Margoliash, 1998; Trifa, 2006; Jančovič and Köküer, 2015).

The last method that we survey here is of particular interest as it can be part of the feature creation (Lasseck, 2015a) or a recognition method in its own right. Spectrogram cross-correlation takes a segment of the spectrogram and computes the cross-correlation with a set of template calls. It is simple, yet proven to be successful when scanning a specific species with limited call variations (Ulloa et al., 2016; Frommolt and Tauchert, 2014). For example, spectrogram cross-correlation was used to survey screaming piha (*Lipaugus vociferans*) from autonomous field recordings by Ulloa et al. (2016).

Most of the classification methods that have been used have strengths and limitations, as is identified in Table 2.3. However, before any of them can be used for automated processing of field recordings there are three main issues that need to be considered: noise, scalability, and the addition of new bird calls. In essence, most of the methods are trained and tested on relatively low numbers of high quality recordings, whereas field recordings are inherently noisy, and can contain bird calls from many different birds. The theory seems to be that by using high quality data for training, the algorithms will deal well with noise in true recordings, although this is not tested, and does not seem particularly likely. Even when noise is included, it is sometimes added to clean recordings (Zhang and Li, 2015; Jančovič and Köküer, 2011; Ren et al., 2008; Chen and Maher, 2006), and thus unlikely to be typical of real environmental noise. While denoising methods (as were discussed earlier) can help by improving the quality of the call, this needs to be used for the training data as well as the test data to ensure that the methods deal well with any artefacts that the denoising introduces. In an interesting approach, Papadopoulos et al. (2015) attempted to overcome the non-bird sounds associated with bird recordings by developing individual models for each target species (in their case, 15 species) using novelty detection, so that the model discriminates the target species from noise. Using audio with SNR of -3–3 dB (short length/manual), their AUC ranged from 0.56 to 0.90; high variability in AUC between the species was observed.

Most methods perform progressively less well as more calls from more species are introduced. This is relatively difficult to overcome. However, it is possible to improve the quality of recognition markedly by including information about where a call was recorded, since this can reduce the number of possible species that a call could come

from, given the limited environmental niches of many birds.

In addition, most machine learning methods are trained off-line, based on the complete dataset, before any recognition occurs. This means that if the user wishes to add new calls or species, or even just further examples of calls that are already in the dataset, a new model has to be trained from scratch, potentially a very computationally expensive operation.

2.3 Current Birdsong Recognition Related Software

This section presents an analysis of the software tools currently available, highlighting their strengths and weaknesses in relation to birdsong recognition. A comparison of the various algorithms is given in Table 2.4.

SoundID is a PC-based commercial non-voice sound recognition system that is dedicated to bio-acoustic applications such as animal surveys. It originated in 2003 with the concept of developing an approach to automatically detect a rare parrot species in Australia. The main building blocks are LPC for call pattern representation and the geometric distance (which was developed for the software) to perform the recognition (Jinnai et al., 2012). Recordings are segmented using an energy threshold, and the LPC spectral image computed for each extracted segment. Then the LPC pattern (image) is compared with the stored reference patterns using the geometric distance. Based on our experience, noisy references lead to a decrease in the performance of the recogniser, and the reference calls should be selected very carefully so that they cover all the call variations of the species. At least 20 references are required even for a species that does not display many variations (Boucher, 2014). Understanding of digital signal processing and extensive configuration in each sub-process is essential to gain success from the system. The SoundID group reported more than 95% accuracy in analysing the dawn chorus (Jinnai et al., 2012) and they particularly highlight the efficacy of the software for processing large datasets. However, they are clearly expert users, and in our experience it is hard to achieve a reasonable recall rate and optimisation of the parameters is time-consuming and difficult.

Raven Pro is a stand-alone software program developed by the Cornell Lab of Ornithology for acquisition, visualisation, measurement, and analysis of sounds (Charif et al., 2010). Audio files can be analysed manually, semi-automatically or automatically using the software, which is popular among ecologists as a spectrogram analysis tool (Arévalo and Araya-Salas, 2013; Aland and Hoskin, 2013; Sandoval and Barrantes, 2012; Bura et al., 2011; Kirschel et al., 2011; Crothers et al., 2011; Vernaleo and Dooling, 2011; Aleixandre et al., 2013). The relevant tools provided are noise filtering and manual or semi-automatic syllable segmentation (Stowell and Plumbley, 2011). A couple of syllable-level bird sound detection methods are included that are based on either

estimation of background noise or a user-defined Signal to Noise Ratio (SNR), and there are many other user-defined parameters such as duration, or low-pass smoothing (Charif et al., 2010; Duan et al., 2013). A few studies have attempted to apply these automatic detectors on real world data. For example, Sebastián-González et al. (2015) used band-limited energy detector to detect call events of Hawai'i 'amakihi (*Hemignathus virens*) from field recordings with 93% recall but only 16.8% of them were good selections (precise endpoints). Duan et al. (2013) configured the segmentation module separately for each of five species found in the dawn chorus (based on five hours of data). The overall accuracy is reported as 43%, far below expectation for such a limited number of species. The segmentation was also tested by Fernandez (2012) to detect songs of marine manatees from long recordings collected with automatic recorders. First, they configured and tested the detector on manual recordings that were manually processed previously. The detector was successful in capturing 80% of manually identified recordings, with one false positive. Unfortunately, nothing was gained after applying the detector to autonomous recordings: it only detected 12 calls out of 100 files and all were false positives that were caused by human voices in the marina. An open-source program by the Cornell Lab of Ornithology, eXtensible BioAcoustic Tool (XBAT), is designed to extend its capabilities by allowing user-defined program codes as plug-ins. The developers tested the data template detector on two species, the cerulean warbler (*Dendroica cerulean*) and the whip-poor-will (*Caprimulgus vociferus*). Their best reported results are 100% precision/54% recall and 96% precision/80% recall respectively (Clark and Fristrup, 2009).

Song Scope, developed by Wildlife Acoustics, is also equipped with a call detector (Wildlife Acoustics, Inc, 2011). The feature representation and classification algorithms are grounded on well-established speech recognition methodologies: HMM and MFCC (Agranat, 2009; Duan et al., 2011). In contrast to Raven Pro, Song Scope attempts to detect call structures that are composed of multiple syllables. Initially, Song Scope detects the syllables and then clusters them to form calls. While the developers recommend the software for field biologists to analyse long field recordings made by autonomous recording devices, the main drawback is that their approach is very sensitive to noise (Duan et al., 2013). Therefore, post hoc visual scanning was employed by Buxton et al. (2013) and Cragg et al. (2015) to filter false positives generated by the Song Scope detectors tailored to detect different call types of nocturnal sea birds. Duan et al. (2013) optimized Song Scope's detector for the same data set used for Raven Pro. Overall accuracy was reported as 37%, which is less than with Raven Pro (43%). Noisy training data decreased the accuracy of Song Scope more significantly than Raven Pro because the HMM treats noise segments as syllables and models them

into call structures (Duan et al., 2013). Song Scope was used to detect Cory's shearwater (*Calonectris borealis*) in four colonies by Goh (2011). Sixteen calls with minimal background interference were selected carefully as the training data for the recogniser and the Song Scope recognisers were trained only for male Cory's shearwater due to the poor quality of the female calls. The relationship between the Song Scope detection of Cory's shearwater vocalisations and actual number of vocalisation available in the same recordings was positive (assessed using the Pearson correlation test) and was better than XBAT. Recently, Wildlife Acoustics introduced *Kaleidoscope*, an integrated suit of tools for bio-acoustic analysis advancing Song Scope. While the developers claim that the software can generate a set of reports analysing the sound files (demonstrated in training videos), the success is yet to be evaluated on third party data.

Sound Analysis Pro 2011 is free open-source software for recording and analysis of animal vocalisation based on less-complex features extracted from whole birdsong (Tchernichovski et al., 2000; Tchernichovski, 2012). Although they have considered segmentation of songs into syllables and syllable clustering, they report that in this respect the software has limitations. The primary focus is not the analysis of field recordings: the developers recommend the software to be utilised to train animals with playbacks while recording their vocalisation, e.g., throughout the vocal development of a bird to see the tutor-pupil song relationship (Tchernichovski et al., 2000). Daou et al. (2012) used the features generated from Sound Analysis Pro as the input to their software tool which analyses the syllable transitions within the songs of individual birds.

Avisoft-SASLab Pro (Specht, 1993) is another commercial and general purpose sound analysis software that was created by Avisoft Bioacoustics in 1990. The company also provides a bioacoustics recording device and separate recording software. (Frommolt and Tauchert, 2014) successfully used the software to recognise Eurasian bitterns (*Botaurus stellaris*), a species that generates very low frequency and relatively simple vocalisations. The software is useful to automatically measure sound parameters of the spectrogram and waveform. A long list of publications that used this software can be found under the reference list on their web site.

Arbimon is a web-based network for storing, sharing, and analysing acoustic data. Their cyber infrastructure includes a solar-powered remote monitoring station that sends 1-min recordings every 10 min to a base station, which relays the recordings in real-time to the project server, where the recordings are processed and uploaded to the project website (Aide et al., 2013). Recordings to be analysed need be uploaded to Arbimon in order to see the results. However, they do not report the accuracy, and their online recognition facility is expensive and not feasible for the processing of long field recordings.

Table 2.4: A summary of currently available software.

Software	Open Access	Main Purpose	Pre-processing	Segmentation	Feature representation	Recognition	Performance
SoundID	No	Automatically process field recordings and estimate call counts	Band-pass filter	Time domain energy threshold	LPC	Geometric distance	Developers claim the system can challenge a human expert. Our own best efforts to scan field recordings (morepork) achieved only 40% recall (92% precision) after time consuming parameter tuning (Brighten 2015)
monitoR (R package)	Yes	Automatically process field recordings and estimate call counts	Band-pass filter	Detections are based on user-defined threshold either on the score envelope generated by binary point matching or spectrogram cross-correlation	Two options: mapping anticipated regions of signal within a spectrogram (for binary point matching) or matrix of amplitudes (for spectrogram cross-correlation)	Two options to score each time frame of the test signal: binary point matching or pearson correlation – choosing a score threshold is a user-driven process	The developers evaluated the system with two distinctive species considering only one call type from each species – identification accuracies for two call types were 64% and 72% (binary point matching) and 73% and 72% (using spectrogram cross-correlation) – one template from each call type
Raven Pro	No	Manual review of spectrograms and measure sounds	Band-pass filter	Energy threshold in time (amplitude detector) and frequency (band-limited energy detector)	Facilitates to extract a list of time-frequency features	Not addressed	Developers report that the detectors give false positives even when optimally configured

Song Scope	No	Manual review of field recordings	Noise filtering – requires an estimate of noise	Based on the energy threshold (user-defined frequency band)	MFCC	HMM	Noisy data dramatically decrease the accuracy – recommend to extract training examples from a range of recordings
Kaleidoscope Pro	No	Bat call analysis (also supports non-bat sounds)	Band-pass filter – discard noise only files initially (bat)	A threshold guided method	Zero-crossing – duration	HMM	Uneven accuracy across the species
Avisoft-SASLab Pro	No	Spectrogram analysis	Removes noise below a user defined threshold in frequency	Energy thresholding in time domain	Peak frequency, amplitude at peak frequency, bandwidth, and number of harmonics	Facilitates to cross-correlate sounds	Fail to capture faint calls – encourage to use good quality recordings – developers make no claim about the accuracy
Sound Analysis Pro	Yes	Assessment of vocal imitation and song development in birds	Focus on noise-free recordings	Complete songs (algorithm does not require the songs to be partitioned into syllables)	Wiener entropy, spectral continuity, pitch, frequency modulation	Based on Euclidean distance	Enough examples to cover call variations
Arbimon	No	Online storage of short recordings and upload the final analysis	Not disclosed and could not determine	Energy threshold in frequency domain	Frequency range, duration, maximum intensity, and bandwidth	HMM	Developers make no claim about the accuracy – different models for different species (call types) are needed
Praat	Yes	General purpose sound (human speech) analysis	Band-pass filter	Not addressed	Time domain and frequency domain features	Equipped with NN, but not encouraged to use for real-world applications	Useful to annotate and manually segment sound – equipped with programming extension (scripts)

Software	Open Access	Main Purpose	Pre-processing	Segmentation	Feature representation	Recognition	Performance
Sonobat	No	Bat call analysis	Filter noisy files and to discard them initially	A threshold guided method	Time-frequency and time-amplitude features (72)	Not disclosed and could not determine	Fails to detect poor quality calls
BatSound	No	Basic bat call analysis	Not addressed	A threshold guided method	Manually measured frequency and temporal features	Not addressed	Far behind the practical requirements
Luscinia	Yes	Archive, measure, and analyse recordings	Band-pass filter	Not relevant or addressed	15 acoustic parameters (contours and hierarchical information)	DTW	>20 publications, none of them used the software for surveying birds except behavioural studies based on their acoustics
Syrinx	No	Playback of animal sounds, recording, and analysis	Noise profiling – band-pass filter	A threshold guided method	Time-frequency measurements	Not addressed	Useful for ecologists to playback with a minimum of equipment
PAMGuard	Yes	Passive acoustic monitoring – ocean acoustics	Median filter – average subtraction – thresholding – Gaussian smoothing	Energy threshold	Time-frequency measurements	User-defined classifiers	Not reliable with more species (accuracy dropped from 95% to 59% when the number of species changed from 4 to 12)
SongSeq	Yes	Syllable clustering	Not addressed	Not relevant or addressed	Sound Analysis Pro generated features	Not relevant or addressed	Templates are required

SIGNAL	No	Event detector and analyser modules auto extract and measure sound events from recordings	Band-pass filter	Energy thresholding in time domain	Call duration, call rate, peak frequency, frequency range	Method is not disclosed and could not determine	Set of spectral, temporal, and amplitude parameters to be determined by the user
Ishmael	Yes	Manual review of spectrograms	Not disclosed and could not determine	Energy threshold	Not disclosed and could not determine	Spectrogram correlation	Not reported
SpectraPRO, *SpectraLAB*	No	Manual review of spectrograms – focus on ocean acoustics	Not relevant or addressed	Not relevant or addressed	Not relevant or addressed	Not relevant or addressed	Not relevant or addressed

While we discussed some of the software tools above, more are given in Table 2.4. All the systems that we reviewed have strengths and weaknesses on their own. Overall, the software tools are far behind the practical requirements demonstrating that further developments are essential, particularly, to overcome the problems generated by unwanted sounds (noise) in field recordings and also to deal with faint bird sounds recorded due to the spatial distribution of birds in their natural habitats.

2.4 Conclusion and Recommendations

Evaluations of recordings from programmable acoustic recorders on field observation-based call count surveys suggest that unattended recordings have the potential to establish the presence of bird species and to replace costly manual call count surveys (Venier et al., 2012; Haselmayer and Quinn, 2000; Hobson et al., 2002; Acevedo and Villanueva-Rivera, 2006; Alquezar and Machado, 2015; Klingbeil and Willig, 2015; Borker et al., 2015; Sedláček et al., 2015; Zwart et al., 2014; Towsey et al., 2014; Furnas and Callas, 2015; Holmes et al., 2014). Recording forest sounds is easy with the available programmable acoustic recorders. However, automated processing of these field recordings for birdsong recognition still remains a challenge and is largely performed manually by means of spectrogram reading and/or listening to the recordings. There is currently a dire need for an automated system that is capable of detecting, recognising, analysing, and inferring bird populations from sound recorders (Bardeli et al., 2010; Potamitis, 2014; Swiston and Mennill, 2009). Such systems will enable the use of many recorders across an environment, for long periods of time, something that is necessary to understand bird populations better. For example, Figueira et al. (2015) used sound recorders to reveal the higher use of Amazon old forest than the secondary forest by parrots. To do this they manually analysed more than 2,000 hours of recordings.

In this paper we have investigated the state of the art of automated bird song recognition research as part of our own on-going research aimed at bird conservation through machine recognition of birdsong. We have identified a major limitation of current research: the inability to effectively recognise bird species from natural unattended recordings. Even though the reported accuracies are impressive, this is because they are largely based on small datasets and less noisy data. For example, Frommolt and Tauchert (2014) detected Eurasian bitterns in controlled field recordings (data collected under calm conditions) with 85% recall. But under real conditions bittern booms are usually hidden under low frequency noises such as wind, thunder, and vehicles passing. The literature has provided enough evidence to confirm that once we have good quality recordings, even standard methods in machine learning are capable of discriminating between small numbers of different species successfully. Hence the underlying problems are reduced to determining which parts of the given long noisy recording

carries birdsong, removing the noise, and then establishing methods that are capable of maintaining a high recall rate as the number of species increases. The alternative is to establish species-specific methods for particular species of particular interest. This review has convinced us the methods that have been successful to screen field recordings are species-specific methods. However the major issue with current species-specific methods is that they cannot be quickly and easily modified to screen a different species than the species in the original study. For example, the feature extraction can be dependent on the species being considered (Digby et al., 2013). Nevertheless, we suggest that specialised (species-specific) methods for passive acoustic monitoring of particular species is better for conservation purposes, because we require high accuracy in the presence of noisy recordings with calls that can be a long way from the microphone and therefore very quiet.

While there is a lot of research focused on developing general methods for the recognition and classification of short, high-quality recordings of birdsong, relatively few researchers (Bardeli et al., 2010; Ulloa et al., 2016; Ganchev et al., 2015; Digby et al., 2013) have addressed the real need of detecting and extracting calls from long field recordings. In our experience, the proportion of the useful sections (signal) to the length of recordings devoid of any birdsong in long-term continuous acoustic monitoring is exceptionally low, meaning that a call detector alone can save a huge amount of time currently dedicated to manual scanning. At this point one can decide to use machine learning, or hire labour to do the classification.

While it is obviously important that the methods be as accurate as possible, the importance of the recall rate for rare species is worth noting, since for these birds confusing their sound with another (more common) species can lead to over-estimate of their abundance, or miss the fact that they exist at all in the area. Unfortunately, current call detection and segmentation methods generally show poor recall rates and high sensitivity to noise. Our own best efforts using SoundID to scan recordings of morepork (*Ninox novaeseelandiae*), the only extant native owl in New Zealand, achieved only 39% recall rate (with 92% precision) (Brighten, 2015) after a time-consuming trial and error process for software parameter optimisation. This finding highlights another point that is important: overall classification accuracy is not necessarily the most useful measure of utility for a birdsong recogniser; recall and precision provide more useful information.

It is important that benchmark datasets are available, so that different researchers can compare their methods on the same datasets. In fact, there are high-quality comparison datasets readily available in the form of bird recognition challenges. One such bird identification task started and continued from 2014, BirdCLEF (`http://imageclef.org/lifeclef/2016/bird`) is a part of LifeCLEF, which includes fish and

plant identification tasks as well. Currently, their dataset includes two sets of recordings (for training and testing) including 999 bird species from Brazil (501 species in 2014) and built from the Xeno-canto (`http://www.xeno-canto.org/`) collaborative database. The Xeno-canto project supports uploading and downloading sample short length song files (mostly recorded manually) and at the moment of writing this review it covers 9,650 bird species all around the world. NIPS4B (Glotin et al., 2013) is another bird identification challenge that was held in 2013 with 1,000 short-length recordings (nearly 2 hours in total) consisting of 87 sound classes (birds, amphibians) from France, based on an idea from the ICML4B workshop that was part of the ICML2013 conference. NIPS4B data were provided by the BIOTOPE (an ecology consultancy firm that has a collection of wild recordings of birds in Europe). The MLSP 2013 bird classification challenge (Briggs et al., 2013) also presented a dataset including 19 bird species recorded with Song Meter acoustic recorders in an experimental research forest, Oregon. The dataset composed of 645 ten-second long audio recordings. However, all these datasets provided short recordings rather than the long recordings that one could usually record with automated recorders. In addition, they are focused on producing reasonable accuracy on multiple recordings of a large number of bird species. For conservation purposes, it is more important to detect individual species extremely accurately, and the detectors can be specialised since the vast majority of the species included in these datasets will not be present in any given conservation area.

The Cornell Lab of Ornithology also maintains an archive of sample bird calls (`http://macaulaylibrary.org/`), which is one of the world's largest archive of wildlife sounds and videos. Tierstimmenarchiv (`http://www.tierstimmenarchiv.de`) at the Museum für Naturkunde in Berlin, maintains a free bioacoustics inventory of short field recordings along with their metadata (such as the species, the place of recording, and the date) where possible. Further, Ranft (2004) provides a comprehensive summary of major natural sound archives. Ranft estimated that more than 90% of world's bird species had been represented in major sound archives by 2003. However, when it comes to automated recognition, an important but difficult part of a data collection is that some of the recordings need to be annotated and time-stamped precisely, providing the ground truth to train and then evaluate the accuracy of any method tested on the data. The annotation requires expert knowledge and is time consuming (Ganchev et al., 2015). In addition, for obvious reasons, these datasets focus on short segments highlighting clear recordings of individual birds, which is far simpler than the full processing of automated recordings. The RMBL-Robin database (`http://www.seas.ucla.edu/spapl/projects/Bird.html`), an American robin (*Turdus migratorius*) database, manually recorded with SM1 recorders is readily available

from UCLA (Chu and Blumstein, 2011). Considering the demand for automated birdsong recognition, as well as the interest in using programmable recorders for conservation, it is clear that there is a huge volume of recordings already stored around the world, covering many bird species. In addition to the processing of the files to recognise birds, there are also many opportunities for the management of this data and making as much of it as possible available to the wider research community.

In this paper we have considered species recognition. However, in order for birdsong recognition software to be really useful for conservation, more work is required: male-female balance within a bird population is a key factor that determines their stability. Acoustic identification of gender is relatively simple when there is a significant difference between male and female vocalisations, for example in the case of kiwi (*Apteryx*) and Ma'oma'o (*Gymnomyza samoensis*). Further, acoustics have proven to be the most promising technique to identify sex even in the presence of high morphometric overlap (Stirnemann et al., 2015). However, in some species the females are silent and only the males vocalise, e.g., bittern spp.

Another extension of passive acoustic monitoring is to detect the presence/absence of predators, or estimate their populations, or even to track humans in protected areas. For example, introduced mammalian pests have been a major threat to the New Zealand avifauna since human colonisation (Saunders and Norton, 2001; Miskelly et al., 2008). Introduced-predator control and eradication are therefore common management tools with the support of the government and community groups. Acoustic monitoring of pest populations could provide an effective platform to assess the pest management methods reliably.

Chapter 3

The Impact of Environmental Factors in Birdsong Acquisition using Automated Recorders

Abstract

1. The use of automatic acoustic recorders is becoming a principal method to survey birds in their natural habitats. As with any other sound, birdsong degrade in amplitude, frequency and temporal structure as it propagates to the recorder through the environment. Knowing how different birdsong attenuate under different conditions is useful to, for example, develop protocols for deploying acoustic recorders and improve automated detection methods.

2. This paper presents playback and re-capture experiments carried out under different environmental conditions using twenty bird calls from eleven New Zealand bird species in a native forest and an open area, answering five research questions: (1) how does birdsong attenuation differ between forest and open space? (2) what is the relationship between transmission height and birdsong attenuation? (3) how does frequency of birdsong impact the degradation of sound with distance? (4) is birdsong attenuation different during the night compared to the day? and (5) what is the impact of wind on attenuation?

3. The results demonstrate that birdsong transmission was significantly better in the forest than in the open site. During the night the attenuation was minimum in both experimental sites. Transmission height affected the propagation of the songs of many species, particularly the flightless ones. The effect of wind was severe in the open site and attenuated lower frequencies. The reverberations due

to reflective surfaces masked higher frequencies (>8 kHz) in the forest even at moderate distances.

4. The findings presented here can be applied to develop protocols for passive acoustic monitoring. Even though the attenuation can be generalised to frequency bands, the structure of the birdsong is also important. Selecting a reasonable sampling frequency avoids unnecessary data because higher frequencies attenuate more in the forest. Even at moderate distances recorders capture attenuated birdsong, hence automated analysis methods for field recordings need to be able to detect and recognise faint birdsong.

3.1 Introduction

The energy of audio signals reduces as they travel. Thus, the energy of the signal received is always lower than that originally produced. While this acoustic attenuation is relevant to any form of audio processing, it is a particularly important issue in outdoor recordings, where the distances can be long, the sources of noise are significant, and there can be objects between the source and recorder. In addition, the amount of attenuation is frequency dependent, meaning that the characteristic appearance of the signal can change. Given the interest in automatic recordings of birdsong, it seems timely to revisit the question of how birdsong attenuates in natural environments, since the degraded birdsong is what is captured by the recorders, and attenuation makes the analysis of such recordings more difficult than it would otherwise be. In this paper we present the analysis of a set of experiments where we investigated the significance of various factors that could affect how birdsong is attenuated with distance in outdoor environments.

The study of sound produced in natural environments from any source (biophony, geophony, and anthrophony), in any landscape, is termed soundscape ecology. This was seemingly first mentioned by Schafer (1977) and subsequently reshaped, focusing on environmental issues and needs such as the assessment of environmental quality, urban planing and design, and long-term monitoring of the effects of climate changes by other authors (Pijanowski et al., 2011a,b; Farina, 2014; Truax and Barrett, 2011).

While a complete description of atmospheric sound transmission is beyond the scope of this paper, we provide a brief overview of the causes of sound attenuation. For further details, see the standard references for acoustics, such as Kinsler and Frey (1962) and Ingård (1953). There are three principal causes of signal attenuation, namely the spherical spreading out of the signal, absorption of the signal by the atmosphere, and the interaction of the signal with other objects, such as the ground, barriers, variations in air pressure, etc. These causes can be combined additively, but their actual modelling

is less clear, since they depend upon the way that the sound is produced (as a plane wave, or from a point source, or inbetween).

The difficulty in computing these effects analytically for any real world example means that experiments are the most informative way to see the actual effects of acoustic attenuation. This is particularly the case in outdoor environments, where the weather plays a significant role: the effect of wind and ground attenuation are reported as the major sources of sound attenuation when compared to humidity, temperature, fog, and rain (Ingård, 1953; Aylor, 1972). It has been reported that ground attenuation has more influence when the sound source and the receiver are close to the ground (Ingård, 1953). In addition, the environment also plays a role, and this was studied in the late 1970s, with research investigating sound propagation and attenuation with atmospheric transmission, mainly to understand the evolution of acoustic communication and ecological sources of natural selection in birds (Ken et al., 1977; Morton, 1975; Waser and Waser, 1977; Marten and Marler, 1977; Wiley and Richards, 1978; Richards and Wiley, 1980).

Habitat type and recording conditions are assumed to have a strong effect on the quality of the bio-acoustic signals that are recorded with autonomous recorders, and experiments are needed to understand this effect. The aim of this study is to understand how the bird calls recorded degrade with distance in a variety of environmental and weather conditions. This can help in the design of protocols for the use of automatic recorders as well as increasing the accuracy of the analysis of the recorded calls, whether by human or machine.

Our experiment has a simple playback design: a sequence of bird sounds were broadcast from a speaker, and rings of recorders positioned around the speaker captured and stored the sound. We compare the signal-to-noise ratio of the sound files produced by each of the recorders. The purpose of this study is to identify which factors affect the quality of the birdsong that is recorded. The factors tested were (1) open space vs. forest, (2) transmission height (perch height), (3) day vs. night, (4) distance between bird (playback) and recorder, (5) wind direction, and (6) the direction of the bird call in relation to the recorders (Fig. 3.1 (a)).

3.2 Materials and Methods

3.2.1 Study Species

We selected a wide range of bird sounds from very low frequency to high frequency, and with varying complexity. A total of 20 different calls/song segments were selected from eleven New Zealand bird species to playback. These are illustrated in Table 3.1. Sound segments were selected using close-range recordings with minimal noise, and

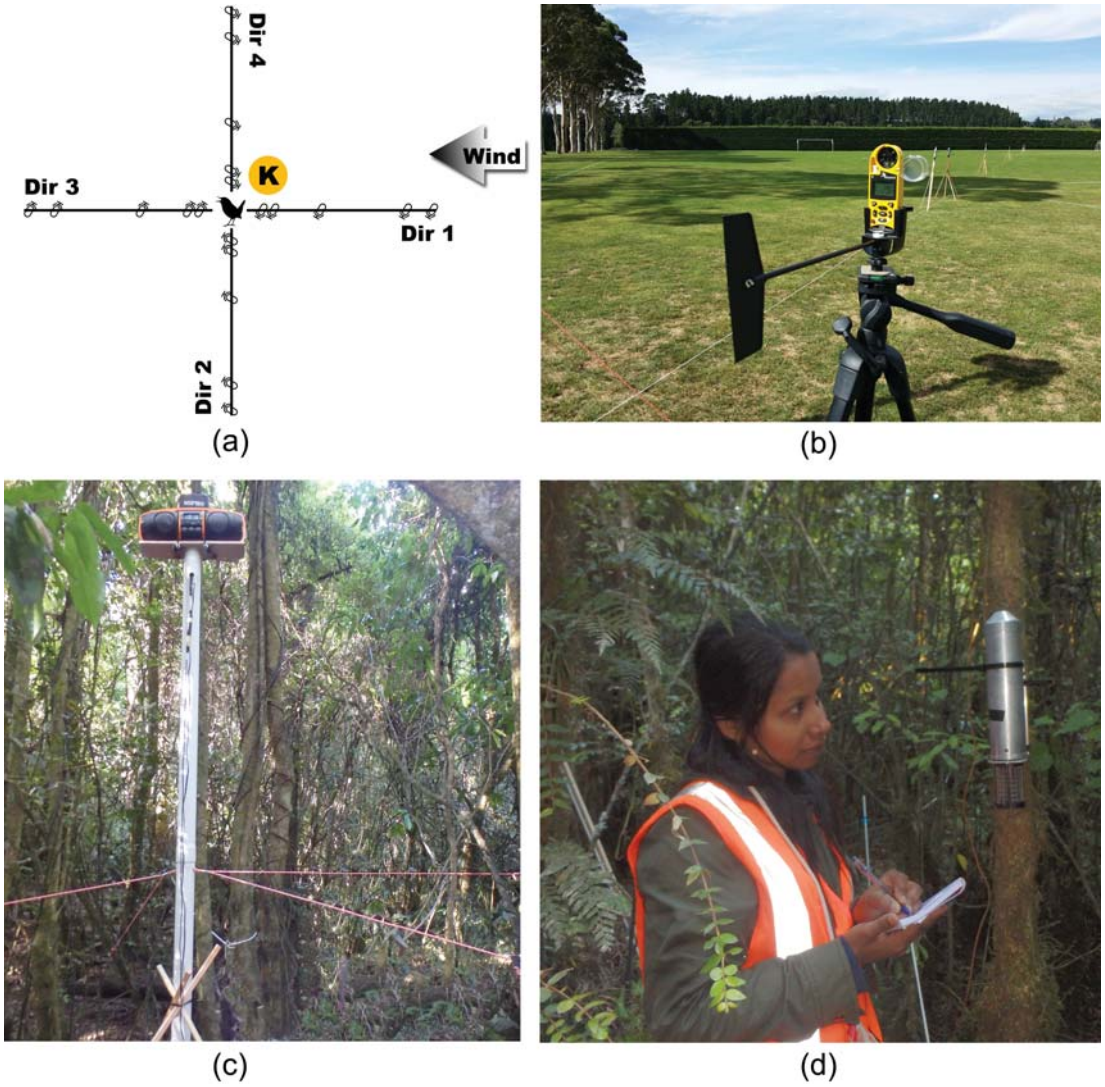
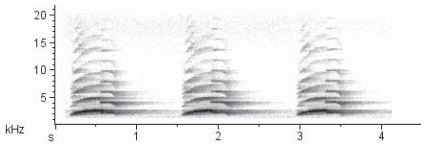
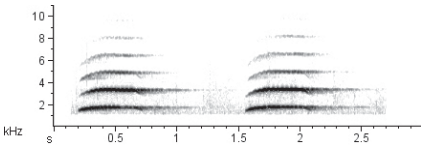
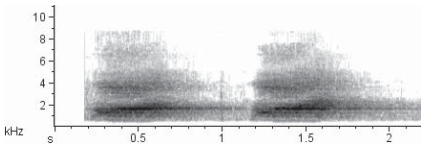
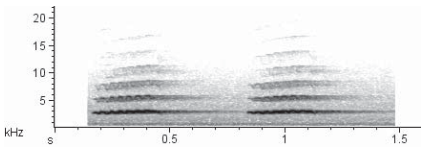
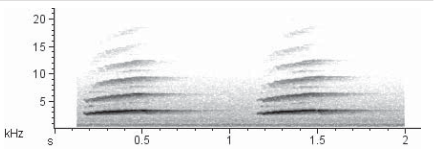
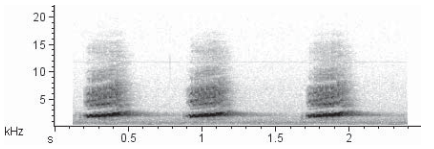
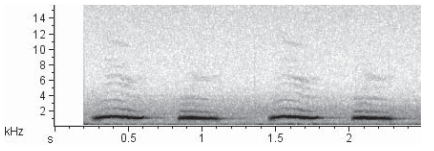
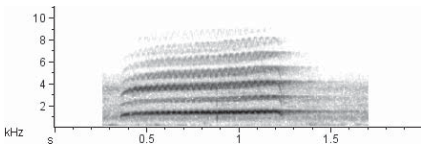
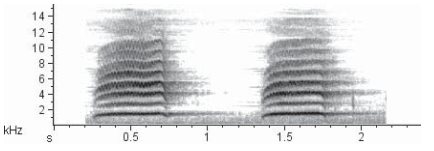
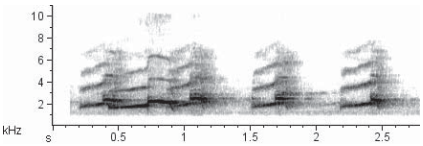
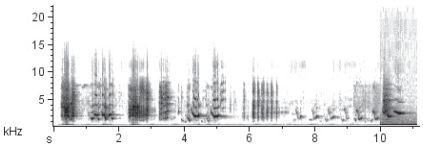
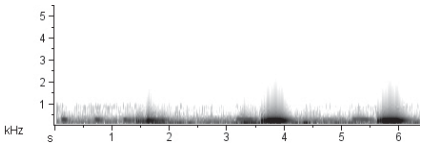
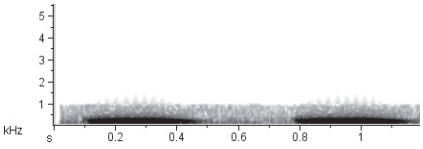
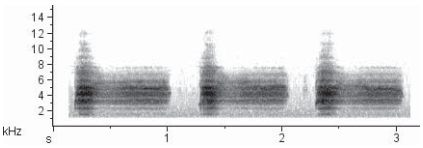
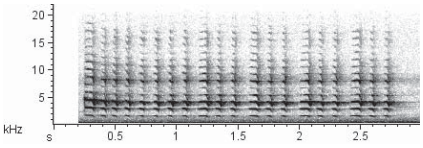
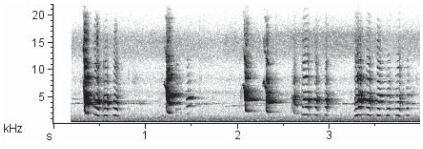
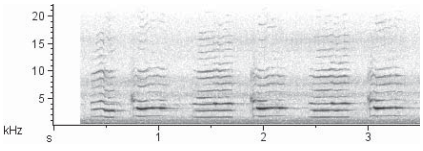
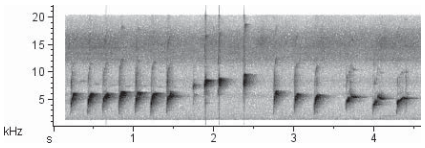
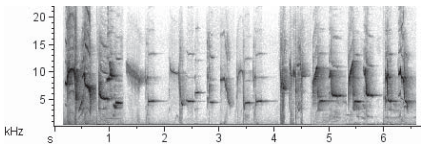
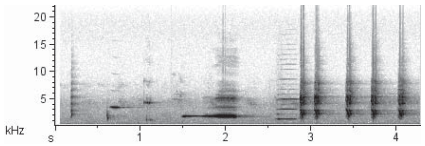


Figure 3.1: (a) Recorders and playback setup, (b) the Kestrel wind meter and a half line of recorders (5 recorders) in open habitat, (c) speaker setup in the forest habitat, and (d) recorders mounted at eye level.

Table 3.1: The bird sounds used in the experiment.

Species	Bird group	Time active	Call type	Label	Spectrogram
North Island brown kiwi (*Apteryx owenii*)	Flightless	Nocturnal	Male	bm1	
				bm2	
			Female	bf	
Little spotted kiwi (*Apteryx mantelli*)	Flightless ratite	Nocturnal	Male	lskm1	
				lskm2	
			Female	lskf	
Ruru (*Ninoxi novaesee landiae*)	Owl	Nocturnal	More-pork	mp	
			Trill (low)	trilL	
			Trill (high)	trilH	

Species	Bird group	Time active	Call type	Label	Spectrogram
Weka (*Gallirallus australis*)	Flightless rail	Nocturnal	Male/ female duet	weka	
North Island kākā (*Nestor meridionalis*)	Parrot	Diurnal		kākā	
Australasian bittern (*Botaurus poiciloptilus*)	Wetland bird	Crepuscular	Boom	bittern	
Kākāpō (*Strigops habroptilus*)	Flightless parrot	Nocturnal	Boom	kBoom	
			Chinging	kc	
North Island saddleback (*Philesturnus rufusater*)	Passerine	Diurnal		sad1	
				sad2	
				sad3	

Species	Bird group	Time active	Call type	Label	Spectrogram
North Island robin (*Petroica longipes*)	Passerine	Diurnal	Male song	robin	
Hihi (*Notiomystis cincta*)	Passerine	Diurnal		hihi	
Tui (*Prosthemadera novaeseelandiae*)	Passerine	Diurnal		tui	

were played back at as close to normal volume for the species as we could judge by ear. The frequency distribution and length of each call (in seconds) can be read off the spectrograms in Table 3.1.

The calls were chosen from several datasets that we had access to; those that were not recorded by the authors are in the acknowledgments. All the calls were captured at 44.1 kHz except the Australasian bittern (8 kHz). The majority of the recordings were made using directional microphones. The bird sounds that had the same volume were concatenated to make a single sound file, as such all 20 bird sounds were fitted into 10 sound files. Each sound file started with a 0.5 sec long acoustic marker (a complex tone) to easily recognise the boundaries of re-captured continuous recordings. Within a sound file the gap between the different sounds was 0.5 sec silence. Accordingly, the total length of playback sounds was approximately 83 sec.

3.2.2 Study Sites

Two relatively flat sites were used to carry out the experiments (Fig. 3.2). The first site was a rugby and soccer field, located at Massey University, divided into four fields using two thin lines of Monterey cypress (*Cupressus macrocarpa*) trees. The distance between the trees was about 3.5m. For the second case, we selected a native New Zealand forest, the Totara reserve (40°7′19.1″S 175°51′17.6″E), in the Pohangina Valley near Palmerston North. The reserve is located beside a river and has a road on the other side of the forest (Fig. 3.2 (b)). The river was almost dry during the course of

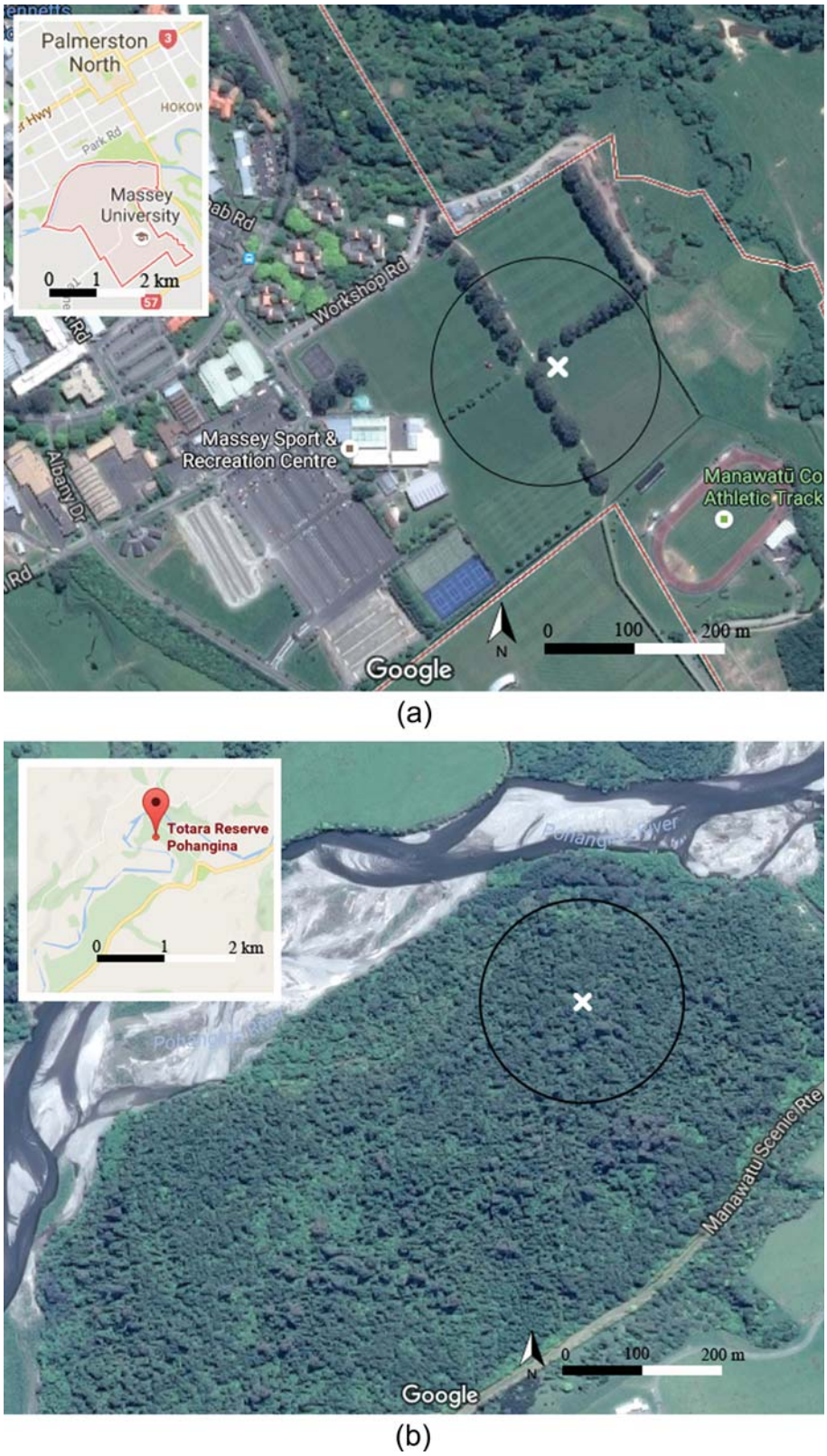


Figure 3.2: The experimental sites: (a) The open site and (b) the forest site. The location of the speaker is indicated by a white cross.

experiments and the recorders did not hear the sound of the river at all. The study site is in the middle of a loop walking track. The selected area is nearly flat and full of large evergreen trees such as totara (*Podocarpus totara*), matai (*Prumnopitys taxifolia*), rimu (*Dacrydium cupressinum*), and kahikatea (*Dacrycarpus dacrydioides*), bushes, and quite a few supplejacks (*Ripogonum scandens*).

3.2.3 Experimental Setup

To observe the effect of the six factors, we set up a playback and re-capture experiment with a single sound generator and multiple recorders. 20 recorders were positioned in 5 rings around the speaker. The rings were located at 20m, 25m, 50m, 100m, and 120m and the recorders were placed at 0°, 90°, 180°, and 270° with respect to the prevailing wind direction. Effectively, the recorders were positioned along two orthogonal lines that crossed at the speaker location, one of which ran towards and away from the wind direction, and one perpendicular to it (Fig. 3.1 (a)). The choice of 20m was made based on preliminary testing, and was sufficiently far away to avoid sound clipping and distortion. The three following distances were simply doubles of each other, while 120m was a practical limit enforced by the size of the field-based site. The wind speed and direction were measured using a Kestrel 5500 Weather Meter (with its vane mount; Fig. 3.1 (b)) set up close to the speaker, but with minimal disturbance to the sound transmission (Fig. 3.1 (a)). Although the Kestrel meter recorded other environmental conditions such as humidity and temperature, those were not treated as factors in our experiment.

All 20 recorders were automatic acoustic recording units created by the Department of Conservation Electronics Laboratory, Wellington (electronics@doc.govt.nz), however there were two versions of the same recorder. We matched recorders (initially 42 recorders) with a similar amplitude/frequency response after preliminary playback-recapture tests using pure sounds (a 'click' sound and tonal sounds) generated manually. The recorders were all mounted at a height of 1.5 m on wooden stakes (with the support of pegs) in the open field or on trees in the forest with the support of a metal bracket to hold the recorder, as is shown in Fig. 3.1 (b) and (d) respectively. The height was chosen as the current practice in the field in New Zealand is to mount the recorders at eye level. All the recorders were mounted so that the microphone was facing the centre, which was where the speaker was. The recorders were scheduled to record at 32 kHz (the highest sampling frequency the selected recorders can achieve).

The speaker was placed on a small platform, and we placed the speaker at two heights: 0.25 m, and 3 m above ground level. These were chosen to simulate ground-based birds, and birds sitting low in the canopy. The mounting of the speaker at 3 m can be seen in Fig. 3.1 (c). While it would have been informative to mount the speaker

even higher, this was eventually ruled out for practical reasons. Two speakers were used: a boombox for the very low frequency kākāpō and bittern booms and a MiPro MA-101c for the other calls. The speaker was connected to a Sony PCM-M10 player via a 5m long cable. Preliminary tests were carried out to tune the volume of the Sony PCM player to output the right volume through the speaker for each sound sample.

The sequence of bird calls was played four times, with the speaker facing in four different directions that corresponded to facing each line of recorders. Table 3.2 summaries the number of repetitions that occurred within one trial for each bird sound. We repeated the playbacks in order to check the consistency over the four directions and to test for the effect of wind.

Table 3.2: Summary of the playbacks and re-captures one bird sound produced.

Transmission direction	**Transmission height**	**Number of repetitions**
Dir 1	low	2
	high	2
Dir 2	low	2
	high	2
Dir 3	low	2
	high	2
Dir 4	low	2
	high	2
Total number of playbacks per bird sound		16
Total re-recordings per bird sound within one trial		320 (= 20 recorders × 16)
Total re-recordings per bird sound in Analysis I		1,280 (= 4 trials × 320)
Total re-recordings per bird sound in Analysis II		320
Total re-recordings per bird sound in Analysis III		1,552 (= 5 trials × 320 - 48 missing)

In the open area the experiment could be set up within half an hour by 6 people, but in the forest, it was extremely difficult to line the recorders up along reasonable lines, and this took most of a day. The time required to complete the playbacks for one trial was approximately one hour including the time to rotate the speaker into the four directions, change the transmission height, switch between the speakers, to adjust the volume, and also to skip some evident background noises such as the calling of wild morepork actually present in the background (who responded to our playbacks of morepork) and sometimes to avoid obviously loud noises made by aeroplanes or vehicles passing. Otherwise, the total length of playbacks was approximately 22 minutes (20 minutes bird sounds and 2 minutes acoustic markers plus silence between the different bird sounds).

The Kestrel meter did not detect any wind in the middle of the forest (close to the speaker) under the canopy, the actual wind during the trials was always around 10

km/h. During one trial (ONM: Table 3.3) three recorders ran out of battery, resulting in 48 missing data points.

Table 3.3: Summary of the trials carried out. The first column consists of the names given to the trials where the first letter corresponds to the location (Open or Forest), the second the time of day (Day or Night), and the third the wind speed (Calm, Moderate, or Windy).

Trial	Open/ Forest	Day/Night	Wind observed by Kestrel meter	
			Median (Km/h)	Range (Km/h)
ODC	Open	Day	3.0 (calm)	0.0-10.0
ODM	Open	Day	7.8 (moderate)	1.2–15.6
ODW	Open	Day	17.5 (windy)	8.4–27.8
ONM	Open	Night	6.6 (moderate)	0.0–14.8
ONC	Open	Night	3.2 (calm)	1.9–7.8
FDC	Forest	Day	0.00 (calm)	0.0–0.0
FNC	Forest	Night	0.00 (calm)	0.0–0.0

3.2.4 Wind Direction

The lines along which the recorders were placed were positioned according to the predominant wind direction at the start of the experiment according to the Kestrel meter (and compared to the weather forecast). However, the wind direction was not constant, and varied significantly during some trials. Therefore we referenced the direction of the wind when each call was played, coding them into 45° blocks, where the centre of zone 1 (Fig. 3.3) was the principal wind direction.

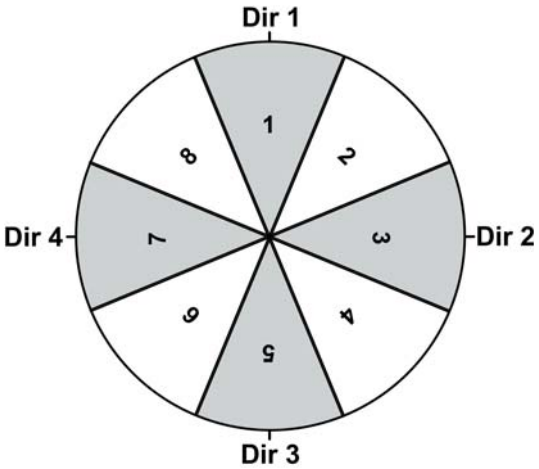


Figure 3.3: The actual wind direction was measured for each bird sound in each trial by post processing Kestral data.

3.2.5 Data Extraction from Continuous Recordings

Once a recorder was started (scheduled), it continued recording until we stopped it at the end of the experiment/until the end of the schedule. We used a set of acoustic markers to identify when each experiment started, when the transmission height was changed, or the speaker direction changed. These markers enabled us to precisely time stamp the recordings, which was particularly important for the recorders that were further away, and did not detect the birdcalls perfectly. We used the software Praat to annotate the recordings, by manually identifying the acoustic markers in one recording (a 20m distant one), and then matching the annotation (TextGrid in Praat) to the other 19 recordings for that trial. This resulted in a text grid with 21 tiers (20 bird sounds and the noise component used to measure the dependent variable). All the re-captures of each bird sound (captured by 20 recorders) were segmented, and stored separately, and then the intensity of the signal measured using Praat scripts. This was a semi-automatic method, but still took considerable researcher time.

3.2.6 Dependent Variable and Covariates

There are a variety of possible measurements that can be used to identify degradation of the audio signal. For example, the loss of higher harmonics can be observed as the recorder and player get further apart. Sound level meter, sometimes referred to as Sound Pressure Level meter can be used to measure the intensity of sound at the receiver when the transmitted signals are pure tones (Marten and Marler, 1977), but we transmitted real bird sounds. Therefore, for this paper, we have chosen one simple measure that captures the most important part of acoustic attenuation, namely the Signal-to-Noise Ratio (SNR). This measure is the ratio of the power of the signal recorded divided by the power of the noise. Thus, a large value indicates a clearer signal.

There are two challenges with using this concept, namely that neither the pure signal nor the pure noise are generally known. We could have compared the broadcast signal and the received one, but this would not include noise added by the speaker. It would also require perfect temporal lining-up of the two sounds. We therefore used a variant of the SNR, which we term SnNR:

$$SnNR = \frac{S+N}{N} \tag{3.1}$$

In this equation, $S + N$ is the intensity of the recorded bird sound, and N is the intensity of the background noise at the recorder. To measure N, four 10 second sections that did not contain audio signal (in between the playbacks) were selected and the power in those segments averaged for each recording. The SNR can be transformed into decibel

units by computing $10\log_{10}(SNR)$. However we did not do this because we required positive values for the dependent variable in order to use our chosen statistical analysis method, described next. Through the analysis, SnNR is our dependent variable. The covariates we manipulated were habitat (open/forest), time of the day (day/night), transmission height (low=0.25, high=3m), distance (20m, 25m, 50m, 100m, 120m), recorder direction (Dir1–4), and wind speed (calm, moderate, windy).

3.2.7 Statistical Method

We explored the predictive value of each of the factors on individual bird sounds using Generalised Linear Models (GLM). The observations were independent, therefore the assumption of GLM was satisfied. Prior analysis of the data confirmed that the distribution of data followed a Gamma distribution, being skewed towards larger values. GLM requires some transformation of $X\beta$ (equation 3.2), the linear predictor of covariates, to guarantee the aditivity. The link function, g, defines this relationship between the random component (probability distribution of the response variable) and the systematic component (the explanatory variables in the model).

$$E[Y] = g^{-1}(X\beta) \tag{3.2}$$

A comparison of the log and the power functions showed that the inverse link function was the best fit with the data. In all cases the goodness of fit improved with the inverse link function (Table 3.4). Pearson's chi-square method was used as the scale parameter method (Anderson et al., 2004), with a hybrid of Fisher and Newton-Raphson methods. P value correction (with sequential Sidak) was carried out to avoid Type I errors (inflation of the alpha level) (Abdi, 2007) because we performed multiple tests of mean effect. Both forward and backward selection were employed to find the optimum model, discarding insignificant effects (Table 3.4; see Table A.1 for details of the final models). Even though in some cases the deviance of the final model (Model IV) was slightly larger (compared to the Model III; Table 3.4), we used Model IV as the optimum model because model III had insignificant factors/interactions while model IV had only significant factors/interactions.

For each bird sound, three GLM models were built using different subsets of data (see data sheets S 3.1–S 3.3 in S 3.6). Analysis I comprised the four trials carried out when the wind was calm (ODC, ONC, FDC, and FNC). There were no missing data in this set, therefore the total number of data points per bird sound was 1,280 (Table 3.2). Analysis II used data from the same four trials used in Analysis I, but the speaker direction was fixed, resulting in the data size being reduced to 320 per bird sound. Analysis III data from the trials carried out in different wind speeds in the open field (ODC, ODM, ODW, ONC, and ONM). For each model we looked at the effect of

each factor separately, and the effect of all possible interactions. The statistical analyses were carried out using SPSS® version 22 with 99% confidence interval ($\alpha = 0.01$).

Table 3.4: GLM model development – goodness of fit in each model was measured using the Deviance.[1]

Call example	Deviance			
	Model I (log link)	Model II (inverse link)	Model III (after forward/ backward)	Model IV (after forward/ backward)
bf	6.680	5.939	3.557	3.564
bm1	5.825	5.675	4.064	4.086
bm2	6.227	5.969	3.781	3.796
lskf	6.216	5.970	4.056	4.080
lskm1	5.725	5.512	3.849	3.850
lskm2	5.940	5.815	4.326	4.330
mp	6.042	5.568	3.025	3.032
trilH	6.248	6.033	4.294	4.299
trilL	5.151	5.093	4.426	4.622
bittern	4.266	4.269	4.026	4.048
kBoom	4.598	4.596	4.044	4.172
kc	5.442	5.423	5.001	–
weka	6.741	6.307	3.775	3.788
kākā	5.512	5.387	4.449	4.455
hihi	6.239	6.222	5.891	5.944
robin	6.939	6.879	6.149	6.191
tui	5.443	5.350	4.224	–
sad1	5.259	5.187	3.358	3.384
sad2	5.652	5.602	4.095	4.158
sad3	5.044	4.975	3.283	3.481

[1]Information criteria is smaller-is-better

3.3 Results

3.3.1 Analysis I

Table 3.5: Analysis I: The main effects found in each model (for each bird sound) at α=0.01. Note that this table was generated from 20 individual GLM statistical tests (for each bird sound example). df=degrees of freedom. In grey data for factors that were insignificant.

Bird sound	Model effect	(Intercept)	Day/ Night	Open/ Forest	Height	Recorder Direction	Bird Direction	Distance
	df	1	1	1	1	3	3	4
bf	Wald Chi-Square	258,020	337	1,041	322	43	3	3,446
	p value	0.000	0.000	0.000	0.000	0.000	0.378	0.000
bm1	Wald Chi-Square	286,412	13	258	0	27	7	1,187
	p value	0.000	0.000	0.000	0.481	0.000	0.085	0.000
bm2	Wald Chi-Square	270,971	26	359	10	20	2	1,577
	p value	0.000	0.000	0.000	0.002	0.000	0.513	0.000
lskf	Wald Chi-Square	270,572	35	435	13	12	1	1,383
	p value	0.000	0.000	0.000	0.000	0.008	0.745	0.000
lskm1	Wald Chi-Square	290,832	59	328	0	27	6	1,790
	p value	0.000	0.000	0.000	0.744	0.000	0.091	0.000
lskm2	Wald Chi-Square	279,342	11	222	0	16	4	1,146
	p value	0.000	0.001	0.000	0.632	0.001	0.265	0.000
mp	Wald Chi-Square	278,838	52	930	114	18	0	1,653
	p value	0.000	0.000	0.000	0.000	0.000	0.826	0.000
trilH	Wald Chi-Square	272,516	19	330	22	9	4	1,153
	p value	0.000	0.000	0.000	0.000	0.026	0.224	0.000
trilL	Wald Chi-Square	318,858	0	199	7	5	5	520
	p value	0.000	0.651	0.000	0.010	0.174	0.184	0.000
bittern	Wald Chi-Square	380,904	58	0	2	21	10	200
	p value	0.000	0.000	0.865	0.212	0.000	0.021	0.000
kBoom	Wald Chi-Square	347,271	63	3	2	16	19	125
	p value	0.000	0.000	0.096	0.188	0.001	0.000	0.000

Bird sound	Model effect	(Intercept)	Day/ Night	Open/ Forest	Height	Recorder Direction	Bird Direction	Distance
	df	1	1	1	1	3	3	4
kc	Wald Chi-Square	301,043	0	124	22	5	3	440
	p value	0.000	0.905	0.000	0.000	0.189	0.356	0.000
weka	Wald Chi-Square	253,858	85	499	50	37	7	2,047
	p value	0.000	0.000	0.000	0.000	0.000	0.089	0.000
kākā	Wald Chi-Square	302,111	15	354	0	10	0	917
	p value	0.000	0.000	0.000	0.506	0.020	0.944	0.000
hihi	Wald Chi-Square	262,249	2	104	27	0	2	366
	p value	0.000	0.142	0.000	0.000	0.922	0.606	0.000
robin	Wald Chi-Square	239,984	1	100	6	2	3	617
	p value	0.000	0.259	0.000	0.014	0.560	0.408	0.000
tui	Wald Chi-Square	303,147	0	240	1	12	3	770
	p value	0.000	0.961	0.000	0.248	0.008	0.351	0.000
sad1	Wald Chi-Square	314,182	16	149	0	13	16	910
	p value	0.000	0.000	0.000	0.515	0.004	0.001	0.000
sad2	Wald Chi-Square	294,047	3	115	0.10	20	17	657
	p value	0.000	0.113	0.000	0.751	0.000	0.001	0.000
sad3	Wald Chi-Square	329,656	2	111	6	32	22	930
	p value	0.000	0.120	0.000	0.019	0.000	0.000	0.000

Day vs Night

There was no significant difference in the SnNR between day and night for passerine birds except one call example of saddleback (sad1; Table 3.5). At night, SnNR was higher compared to the day for most of the other bird sounds (Fig. 3.4). It was evident from these results that the sound transmission of nocturnal birds was significantly better during the night compared to the day. Bittern and kākāpō booms consistently followed the opposite pattern (their SnNR was significantly higher during the day).

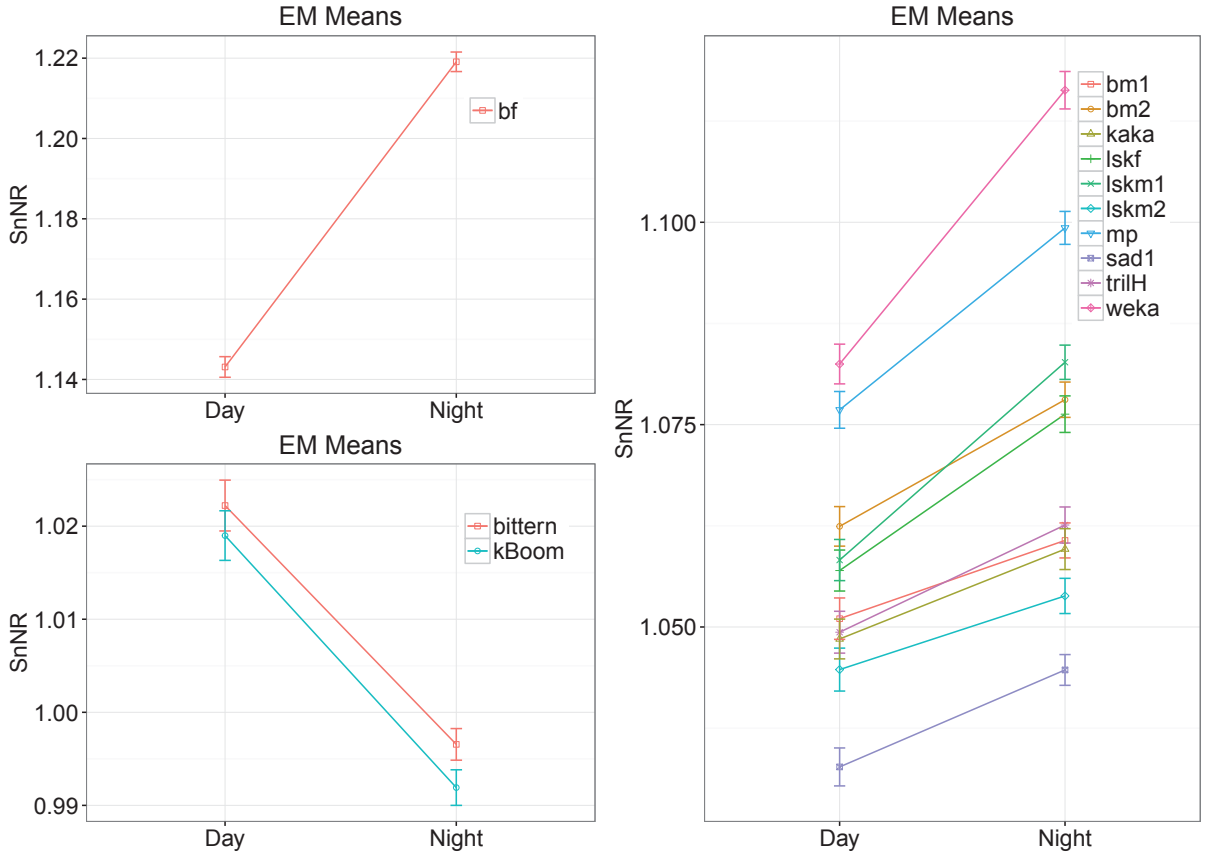


Figure 3.4: Estimated marginal means of SnNR for day vs night. Bars represent standard errors. Note that this figure was generated from 13 individual GLMs (for each bird sound example) and the lines were added to showcase the trend for each test result. bf=brown kiwi female, bm1=brown kiwi male example 1, bm2=brown kiwi male example 2, kBoom=kākāpō boom, lskf=little spotted kiwi female, lskm1=little spotted kiwi male example 1, lskm2=little spotted kiwi male example 2, mp=more-pork sound of morepork, sad1=saddleback example 1, and trilH=trill (high) sound of morepork.

Open vs Forest

In contrast to what we expected, SnNR was always higher in the forest compared to the open area (Fig. 3.5). The only exception was the very low frequency booms of kākāpō and bitterns, which transmitted equally at both sites (Table 3.5).

There was a significant interaction effect of site and the time of day on the two species of kiwis and weka (Fig. 3.6). SnNR was highest when these nocturnal species vocalised in the forest at night and lowest when they vocalise in the open area during the day. The average SnNR for three kiwi examples and weka were similar in the forest despite the time of day, but significantly different from each other in the open site (Fig. 3.6).

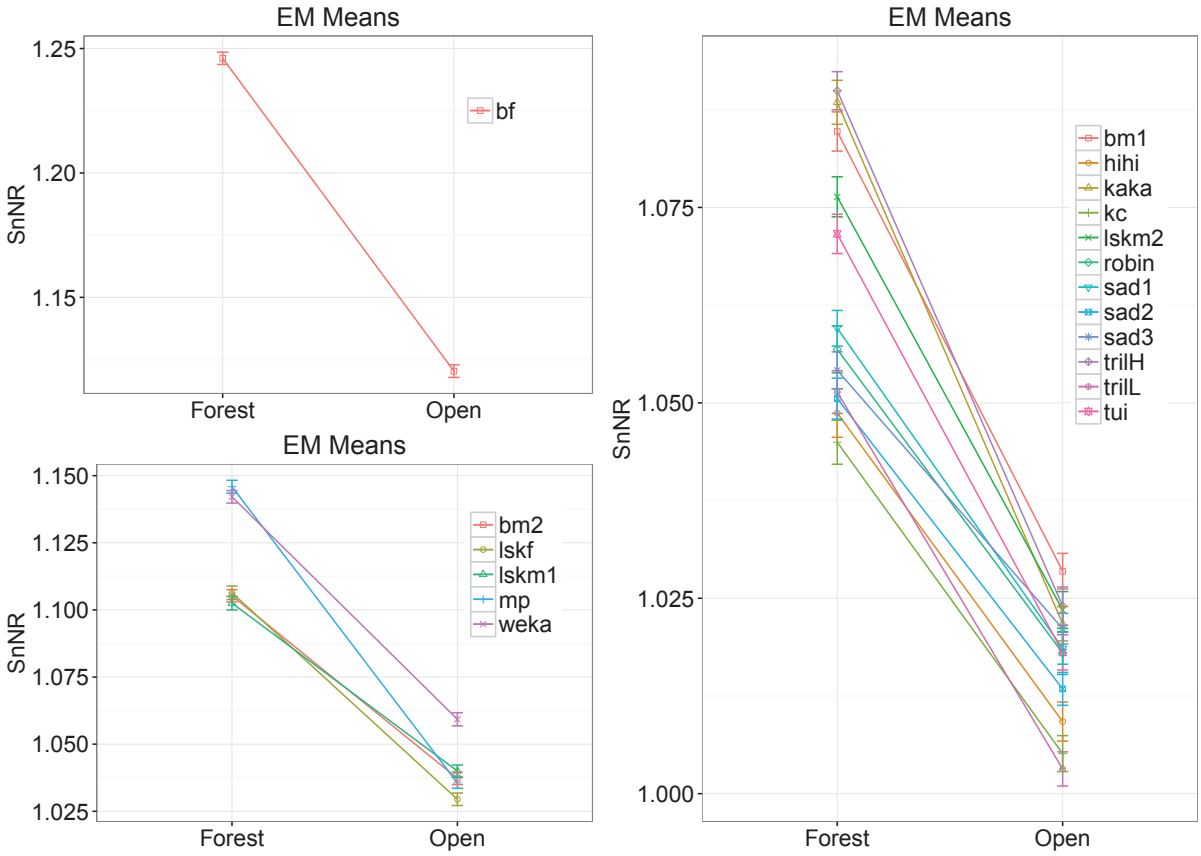


Figure 3.5: Estimated marginal means of SnNR for two different sites. Bars represent standard errors. Note that this figure was generated from 18 individual GLMs (for each bird sound example) and the lines were added to showcase the trend for each test result. bf=brown kiwi female, bm1=brown kiwi male example 1, bm2=brown kiwi male example 2, kBoom=kākāpō boom, kc=kākāpō chinging, lskf=little spotted kiwi female, lskm1=little spotted kiwi male example 1, lskm2=little spotted kiwi male example 2, mp=more-pork sound of morepork, sad1=saddleback example 1, and trilH=trill (high) sound of morepork.

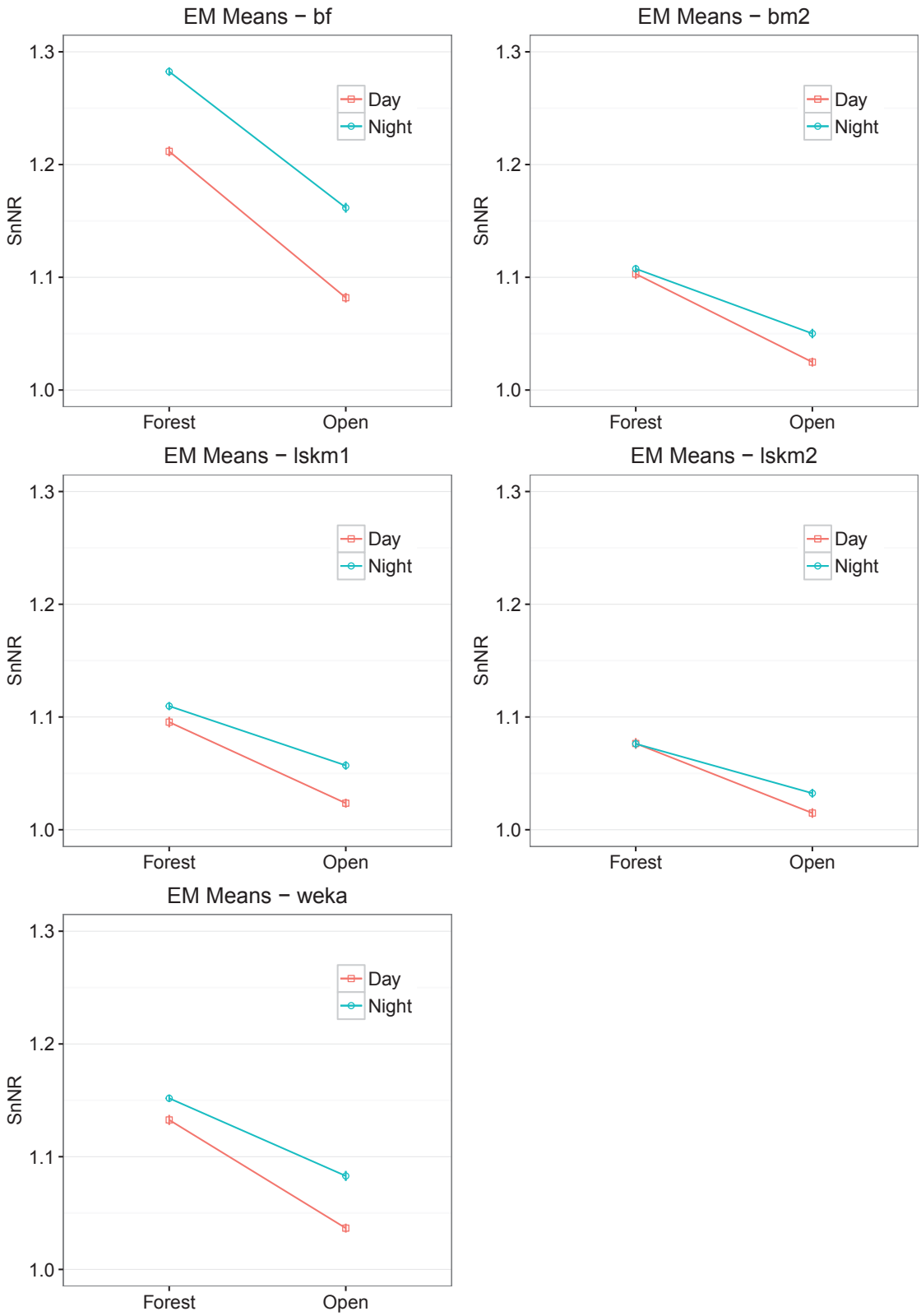


Figure 3.6: Estimated marginal means of SnNR for interaction effect of site and time of the call. Bars represent standard errors. Note that this figure was generated from 5 individual GLMs (for each bird sound example) and the lines were added to showcase the trend for each test result. bf=brown kiwi female, bm2=brown kiwi male example 2, lskm1=little spotted kiwi male example 1, and lskm2=little spotted kiwi male example 2.

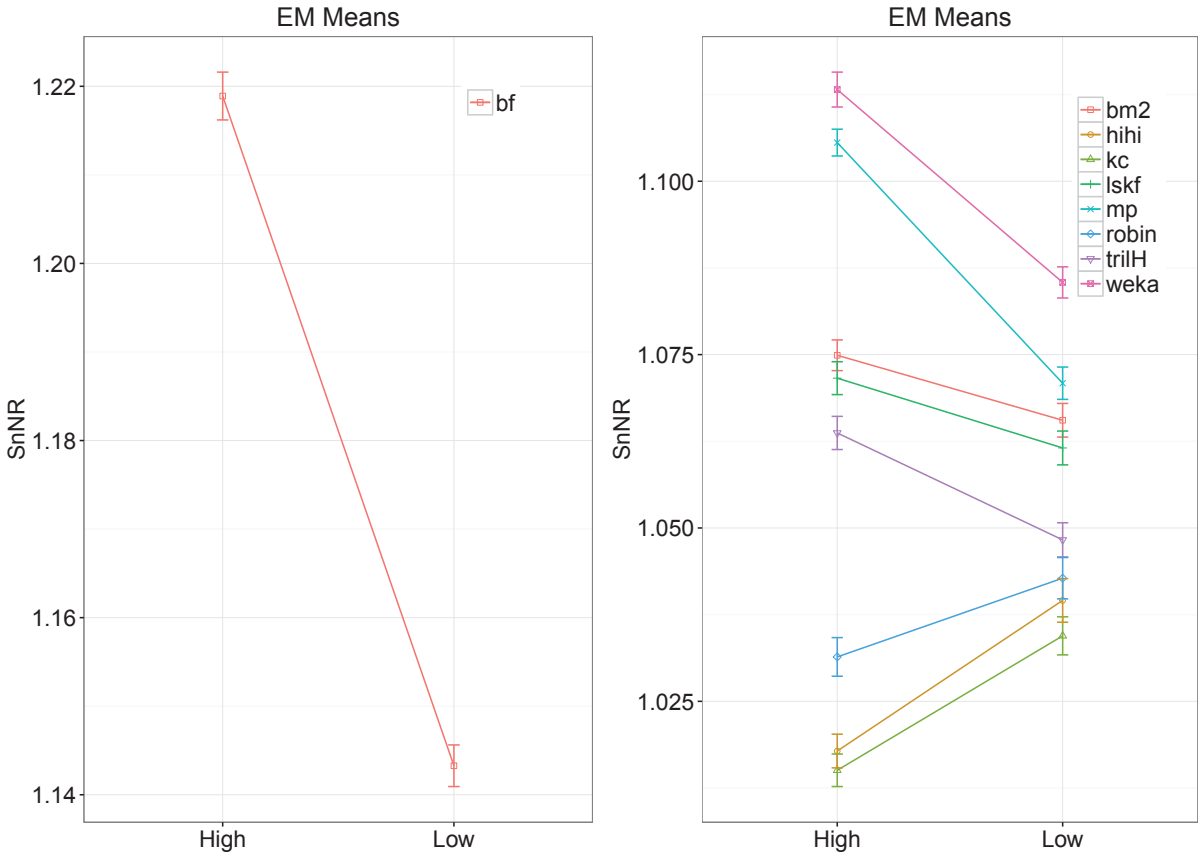


Figure 3.7: Estimated marginal means of SnNR for two transmission heights. Bars represent standard errors. Note that this figure was generated from 9 individual GLMs (for each bird sound example) and the lines were added to showcase the trend for each test result. bf=brown kiwi female, bm2=brown kiwi male example 2, kc=kākāpō chinging, and lskf=little spotted kiwi female, mp=more-pork sound of morepork, and trilH=trill (high) sound of morepork.

Transmission Height

Interestingly the transmission height had a significant effect on some vocalisations of the ground-dwelling species considered (Fig. 3.7). The sound transmission was better at 3m height for two kiwi females and weka. In addition, one sound of the four male kiwi sounds also turned out to be better at 3m height, but the difference was less than with the female call (see Table A.4 for significance). Hihi and robin sounds were better heard when the speaker was close to the ground.

More in line with expectation, the kākāpō chinging sound transmitted better close to the ground (Table A.4) particularly in the open field (Fig. 3.8) and morepork sounds were better heard when broadcast higher. Spectrogram inspection of re-captured more-pork sounds also confirmed that their attenuation was higher when the sound was transmitted close to the ground both in the open site and the forest (Fig. 3.9). The

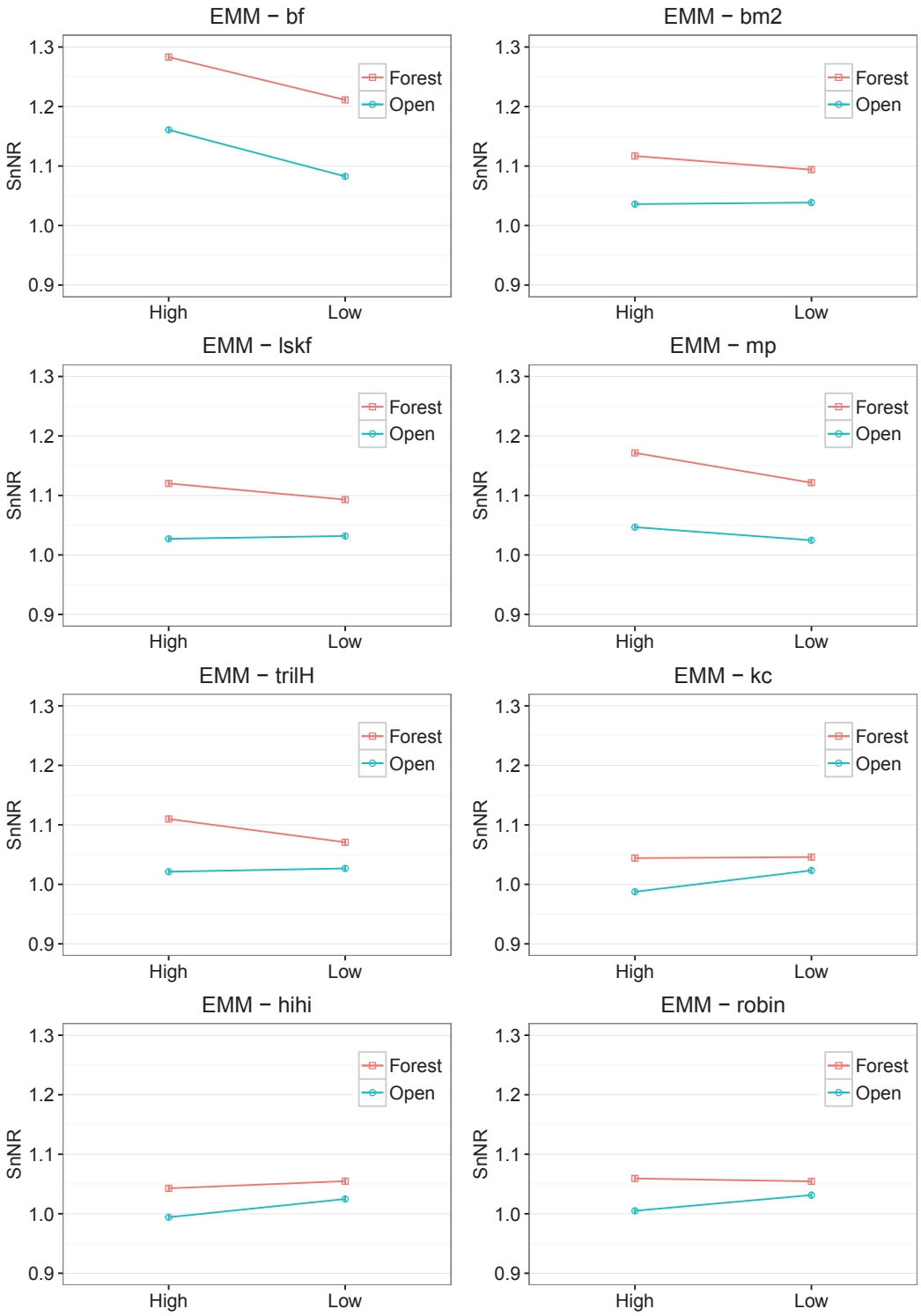


Figure 3.8: Estimated marginal means of SnNR and interaction effect of the transmission height and the habitat. Bars represent standard errors. Note that this figure was generated from 8 individual GLMs (for each bird sound example) and the lines were added to showcase the trend for each test result. bf=brown kiwi female, bm2=brown kiwi male example 2, lskf=little spotted kiwi female, mp=more-pork sound of more-pork, trilH=trill (high) sound of morepork, and kc=kākāpō chinging.

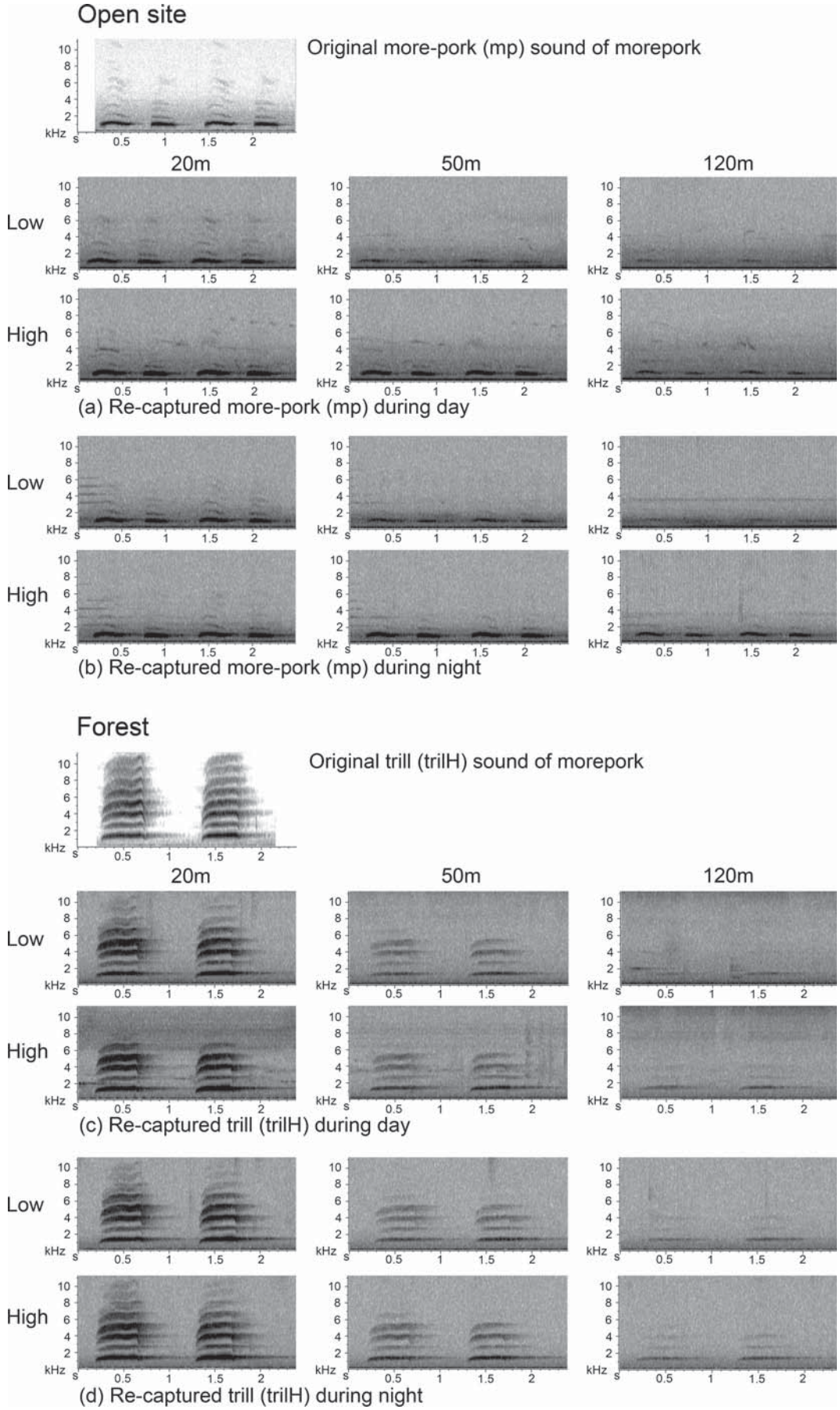


Figure 3.9: Re-captured sounds from morepork broadcasts of more-pork (mp) and trill (trilH). In all cases the speaker was facing the recorders and the wind was calm.

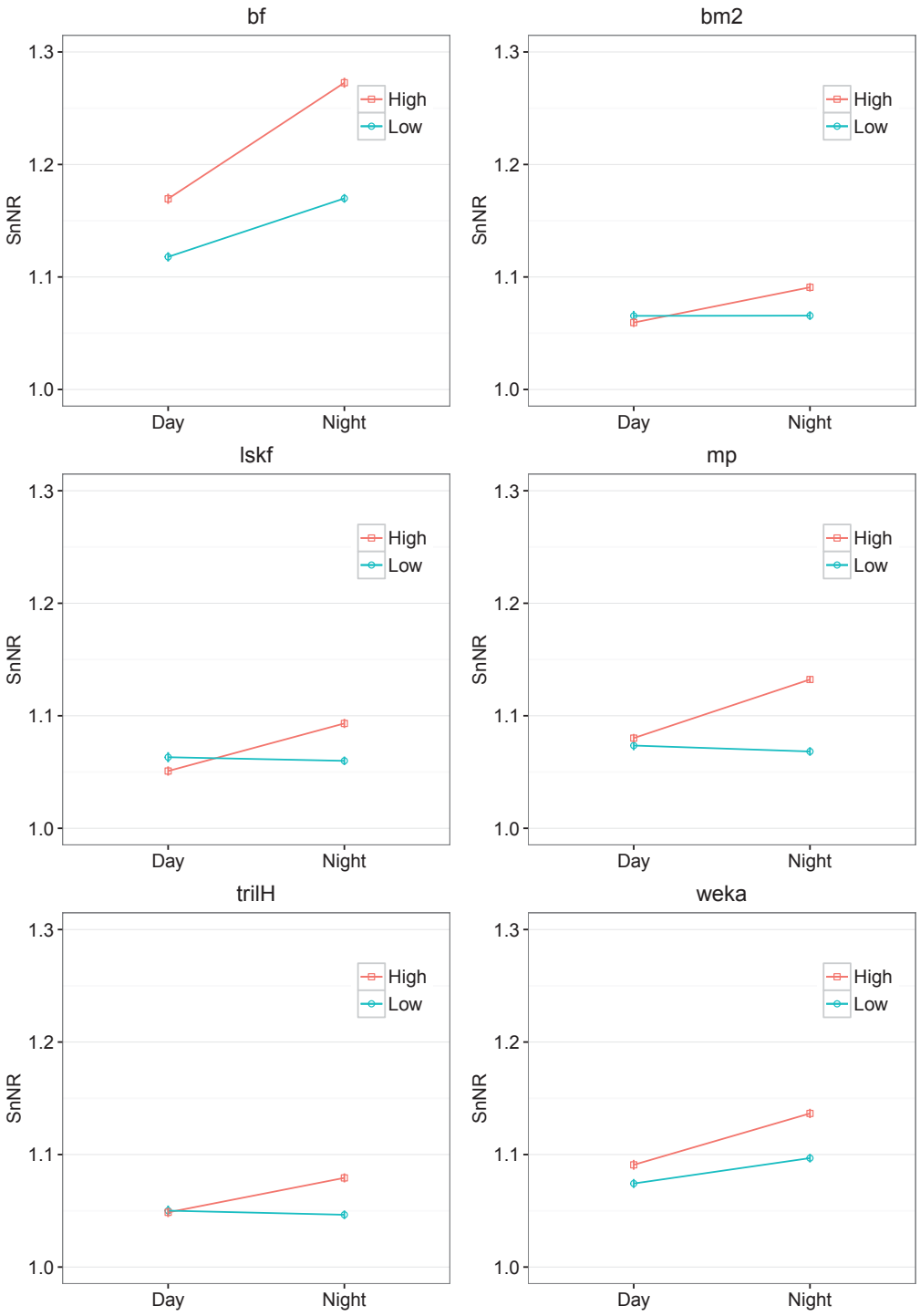


Figure 3.10: Estimated marginal means of SnNR and interaction effect of the transmission height and the time of the call. Bars represent standard errors. Note that this figure was generated from 6 individual GLMs (for each bird sound example) and the lines were added to showcase the trend for each test result. bf=brown kiwi female, bm2=brown kiwi male example 2, lskf=little spotted kiwi female, mp=more-pork sound of morepork, and trilH=trill (high) sound of morepork.

best transmission was always during the night when the sound was broadcast from the 'high' transmission height (Fig. 3.10).

For the sounds of some species there were interaction effects between the transmission height, and the site and the time of the day (Fig. 3.8 and Fig. 3.10). When bird sounds were transmitted at high transmission height, sound transmission was markedly better during the night compared to the day (Fig. 3.10). Overall, high interaction effect between the transmission height and the time of call compared to the interaction effect between the transmission height and the site was evident.

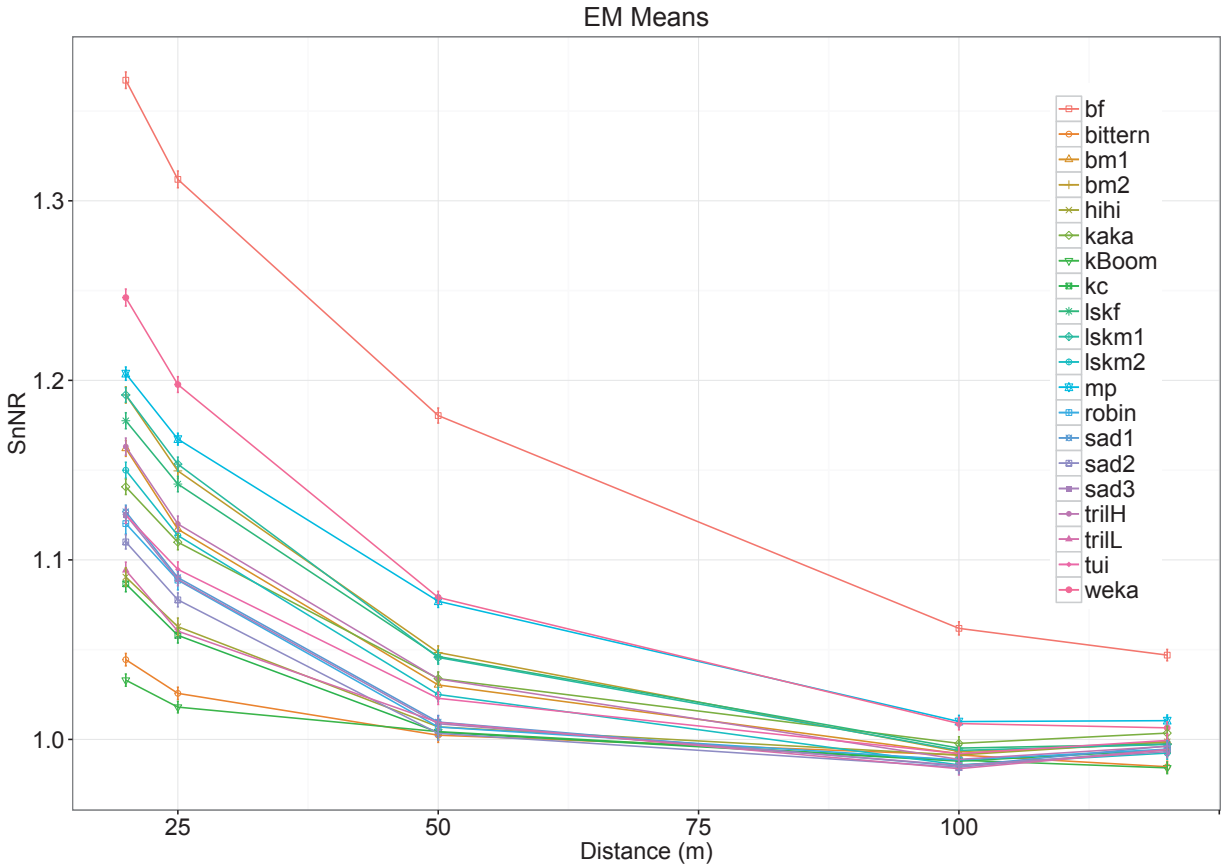


Figure 3.11: Estimated marginal means of SnNR against distance. Bars represent standard errors. Note that this figure was generated from 20 individual GLMs (for each bird sound example) and the lines were added to showcase the trend for each test result. bf=brown kiwi female, bm1=brown kiwi male example 1, bm2=brown kiwi male example 2, kBoom=kākāpō boom, kc=kakapo chinging, lskf=little spotted kiwi female, lskm1=little spotted kiwi male example 1, lskm2=little spotted kiwi male example 2, mp=more-pork sound of morepork, sad1=saddleback example 1, sad2=saddleback example 2, sad3=saddleback example 3, trilH=trill (high) sound of morepork, and trilL=trill (low) sound of morepork.

Distance

As we expected, the experiment confirmed that the SnNR decreases significantly (Table A.5 and A.6) with the distance (Fig. 3.11). SnNR decreased sharply at short distances between the sound source and recorder ($\leq$50) and then slowly at increasing distances.

There was a significant interaction effect (Table A.10) between the habitat and the distance to the broadcast song for all bird sounds except the booms. Recordings in the forest exhibited higher SnNR than those in the open site, the difference was highest at short distance and decreased with increasing distance. The difference was minimal after 100m.

We noted attenuation of birdsong with increasing distance to the recorder in the forest. Even at the relative short distance of 20m, frequencies beyond 6 to 8 kHz were exceptionally attenuated (Fig. 3.9 and Fig. 3.12). However, at 50m calls still carried most of the frequency components they had at 20m but with less energy. The furthest recorder (120 m) did not receive most of the birdsong except for the kiwi and morepork calls.

3.3.2 Analysis II

Analysis II was carried out to investigate the effect of the speaker direction to the quality of the recordings collected. This was intended to reflect the fact that birdcalls are directional. This variable always had a significant effect on the quality (SnNR) of the recordings except in the case of low frequency kākāpō boom (kBoom) and morepork (mp) sounds (Table A.11). As we expected, when the speaker was facing the recorder (Dir1), consistently the SnNR was better (Fig. 3.13 (a)).

Interestingly, the recorders behind the speaker (Dir3) also gained higher SnNR than the recorders in the perpendicular line (Dir2 and Dir4). We presumed that Dir3 had the advantage of wind in the open site. Therefore, we considered the experimental site as a factor. Then it was revealed that wind in the open field had a higher effect than in the closed forest environment and therefore the two open field trials has largely contributed to the shape of Fig. 3.13 (a). For example, Fig. 3.13 (b) illustrates the effect of site for two sound examples, which resulted in a peak at Dir3 (recorders in the line away from the wind direction) compared to Dir2 and Dir4 (perpendicular to Dir3) in the open site.

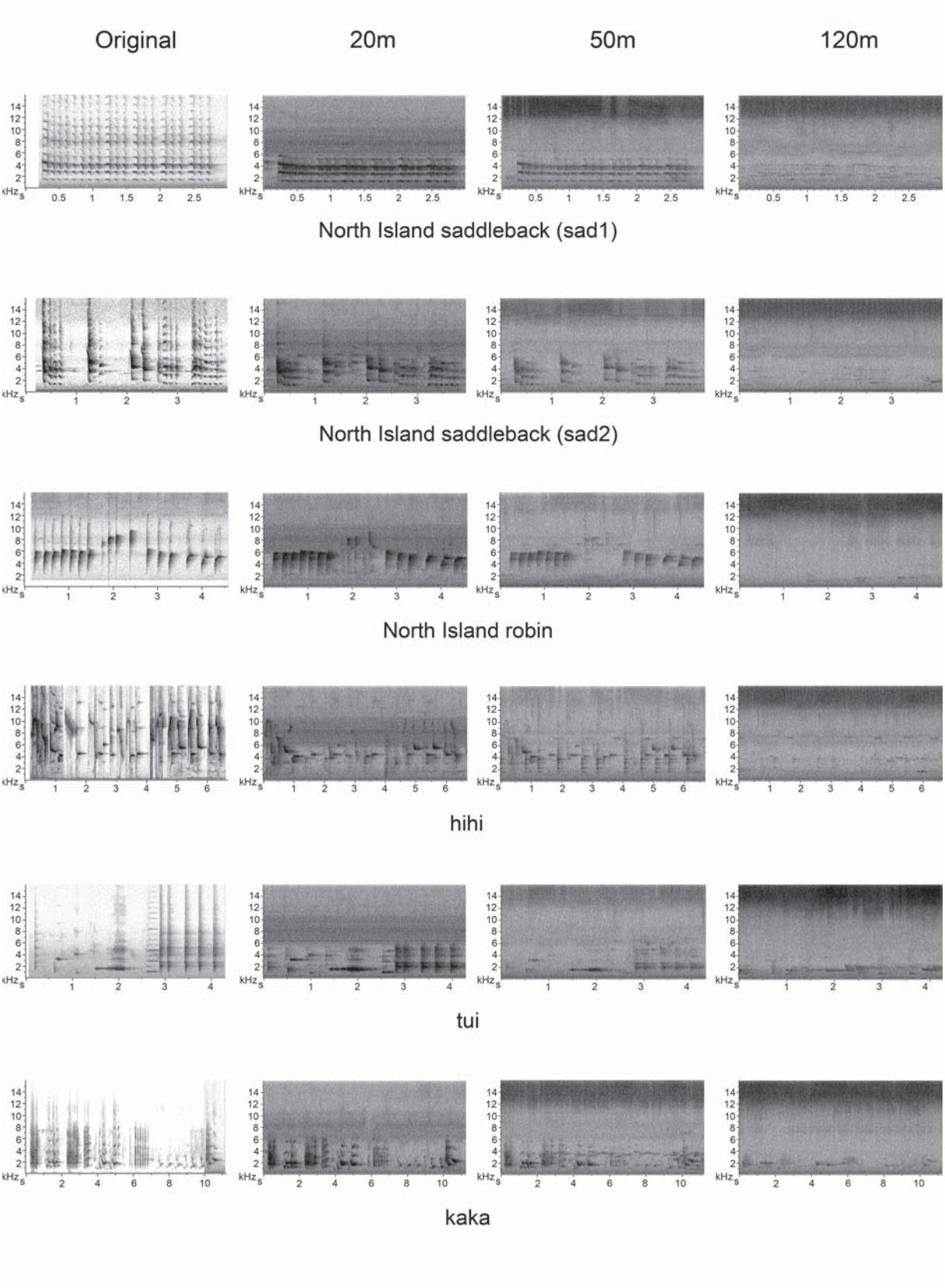


Figure 3.12: Re-captured birdsong that were transmitted from 3m to the ground in the forest during the day. Speaker was facing the recorders (20m, 50m, and 120m) positioned in one direction. (a) North Island saddleback (sad1), (b) North Island saddleback (sad2), (c) North Island robin, (d) hihi, (e) tui, and (f) North Island kākā.

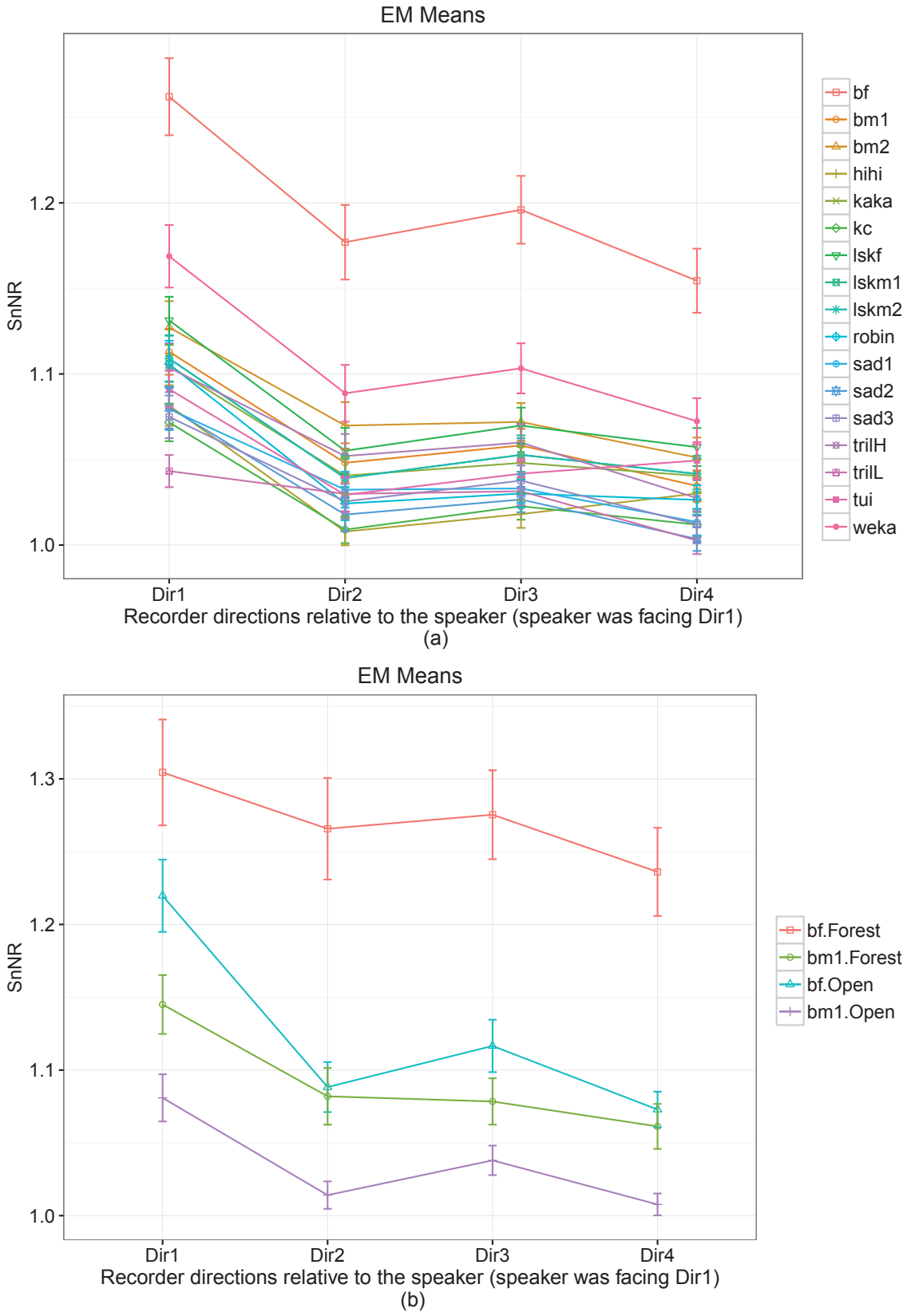


Figure 3.13: The change of the SnNR when the speaker was facing Dir1 (Analysis II). (a) Estimated marginal means of SnNR against the four recorder directions (generated from 18 individual GLMs) and (b) Estimated marginal means of SnNR against the four recorder directions in open vs forest for a male and a female brown kiwi example (bf and bm1; generated from 2 individual GLMs). Dir1–4 are as given in Fig. 3.3. Bars represent standard errors and the lines were added to showcase the trend for each test result.

3.3.3 Analysis III

The influence of wind was more prominent in the open space than the closed forest habitat. Therefore, Analysis III focused on five trials carried out under different wind conditions, 'calm' (<4 km/h), 'moderate' (6–8 km/h), and 'windy' (>15 km/h) in the open site (Table 3.3). As we predicted, the overall highest SnNR was gained by the recorders in the line away from the wind direction (Dir3 in Fig. 3.14; Fig. 3.1 (a)) because the sound waves were guided more to that direction by the wind. The relative direction of the speaker to the recorder was not significant only in the case of kākāpō booming (Table A.12).

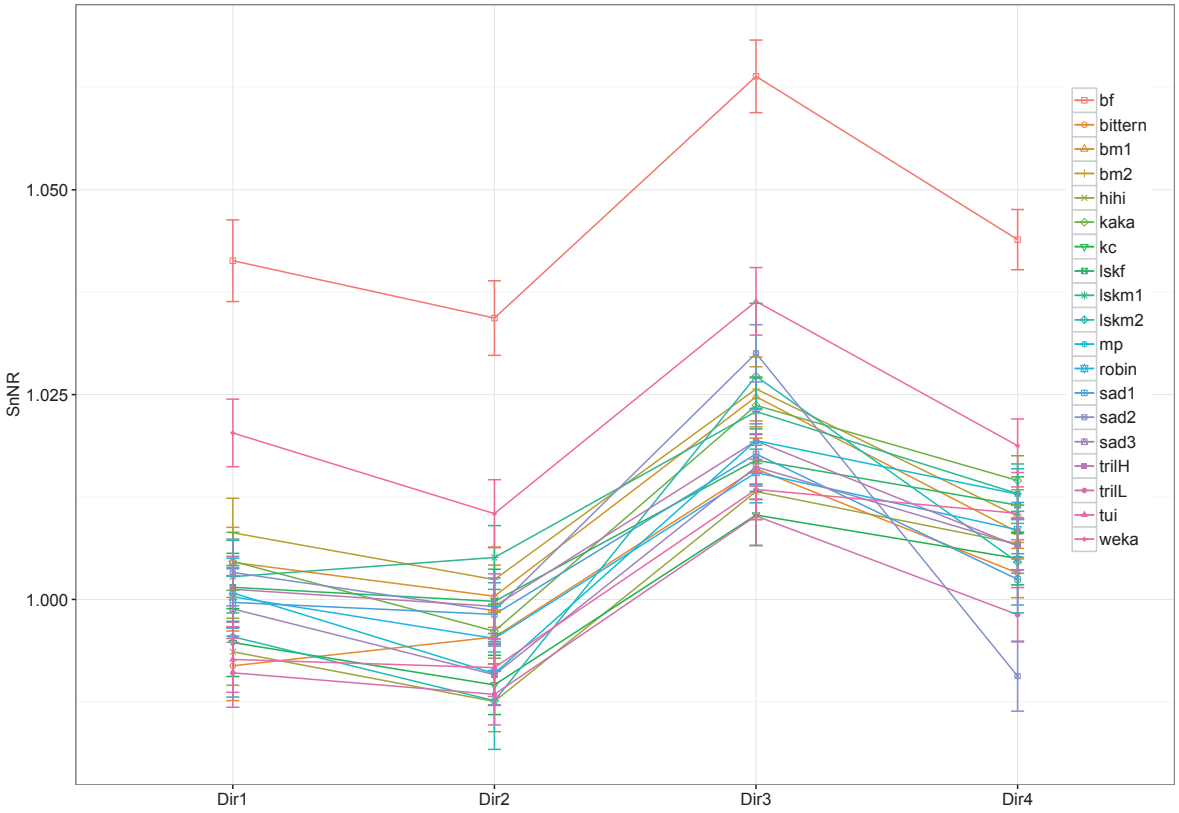


Figure 3.14: Estimated marginal means of SnNR against the four recorder directions when the bird calls to all four directions equally (Analysis III). The dataset includes 'calm', 'moderate', and 'windy' data in the open site. Dir1–4 are as given in Fig. 3.3. The lines were added to showcase the trend, otherwise the lines do not mean anything. bf=brown kiwi female, bm1=brown kiwi male example 1, bm2=brown kiwi male example 2, kc=kākāpō chinging, lskf=little spotted kiwi female, lskm1=little spotted kiwi male example 1, lskm2=little spotted kiwi male example 2, mp=more-pork sound of morepork, sad1=saddleback example 1, sad2=saddleback example 2, sad3=saddleback example 3, trilH=trill (high) sound of morepork, and trilL=trill (low) sound of morepork.

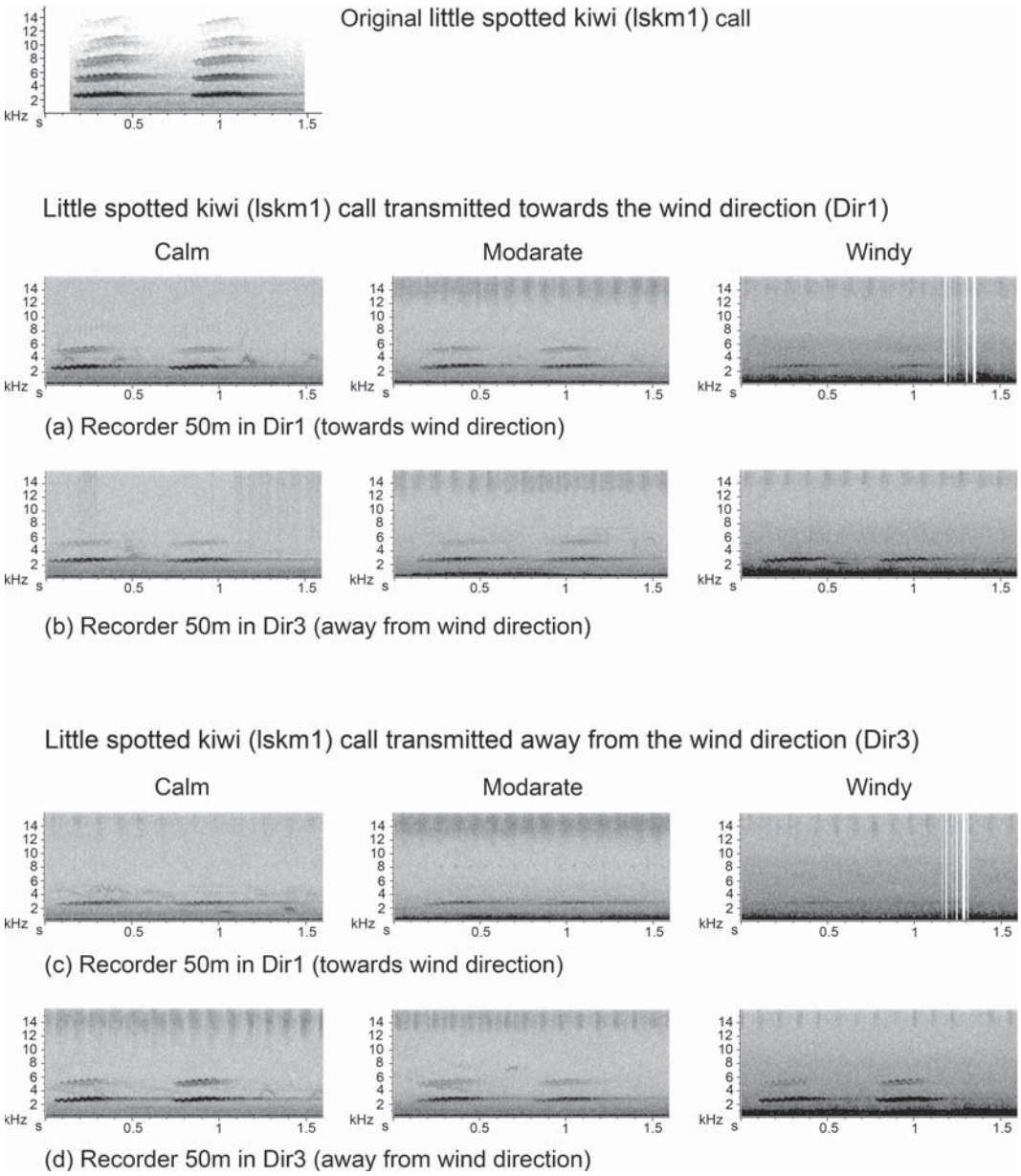


Figure 3.15: Re-captured little spotted kiwi call (lskm1) transmitted close to the ground in the open site under different wind levels when the speaker was facing the wind direction and away from the wind direction. The vertical lines in the spectrograms in first and third rows in the third column are due to gusts of wind while the high frequency noise visible as dark line shadows in most of the spectrograms is possibly due to the noise of a watering machine. Compare to the original sound to discriminate any noise from the bird sound.

The effect of wind is visualised in Fig. 3.15 using one instance of bird sound, a male little spotted kiwi (lskm1). Regardless of speaker direction, the recorder facing away from the wind directon (Dir3) captured the bird sounds better, particularly when the

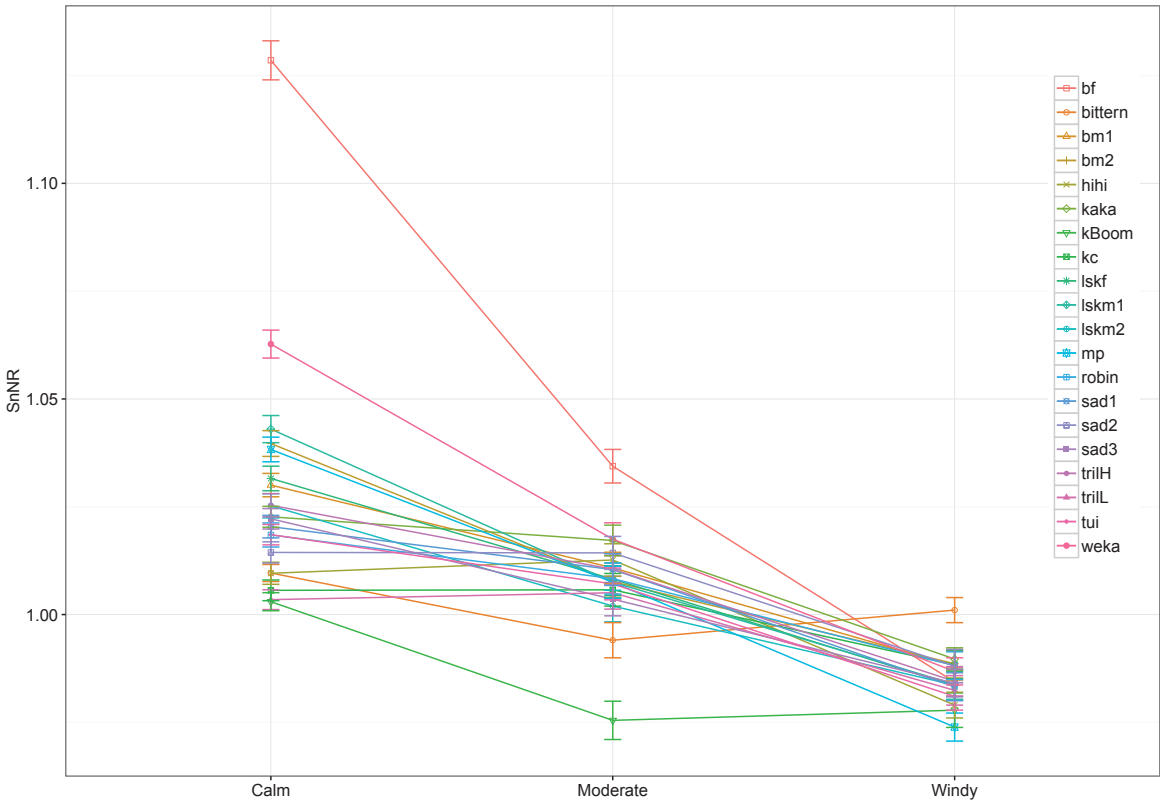


Figure 3.16: Estimated marginal means of SnNR against the different wind levels (Analysis III). The lines were added to showcase the trend, otherwise the lines do not mean anything. bf=brown kiwi female, bm1=brown kiwi male example 1, bm2=brown kiwi male example 2, kBoom=kākāpō boom, kc=kākāpō chinging, lskf=little spotted kiwi female, lskm1=little spotted kiwi male example 1, lskm2=little spotted kiwi male example 2, mp=more-pork sound of morepork, sad1=saddleback example 1, sad2=saddleback example 2, sad3=saddleback example 3, trilH=trill (high) sound of morepork, and trilL=trill (low) sound of morepork.

wind was strong. In contrast, sounds were lost in the spectrograms from the recorder positioned in the direction of the wind (Dir1). The wind intensity significantly reduced the SnNR of the recordings for all the bird sounds (Fig. 3.16; Table A.14). However, kākāpō and bittern booms did not show a significant difference at 'moderate' and 'windy' levels while kākāpō chinging (kc), and the trill sound of morepork (trilL) did not show a significant difference at 'calm' and 'moderate' wind (Tables A.13 and A.14).

3.4 Discussion

3.4.1 Sound Attenuation, Frequency, and Habitat

In normal atmospheric transmission of sound, attenuation is mainly caused by signal spreading, atmospheric absorption, ground attenuation, scattering of sound and deflection by stratified fields (Wiley and Richards, 1978). Natural environments make it extremely difficult to predict sound attenuation reliably; in real outdoor conditions, distance from the source is only one factor amongst many. Amplitude fluctuations and reverberations were studied by broadcasting pure tones by Richards and Wiley (1980) in an experiment similar to ours. They observed that usually higher frequencies attenuate more with the distance and are more vulnerable to both amplitude fluctuations and reverberations. Irregular amplitude fluctuations, mainly caused by atmospheric turbulence from the wind, are more severe in open fields than closed forest habitats and mask low frequencies.

Reverberations are mainly generated by the scattering of sound from reflective surfaces such as tree trunks and foliage surfaces, and are hence more relevant to forest habitats and mask high frequencies (Wiley and Richards, 1978). Spectrograms get blurry and tonal sounds with sharp start and end arrive at the receiver with progressive onset and long reverberations (e.g. Fig. 3.17) because omni-directional recorders (automated recorders) pick up the signals that are scattered and reflected by trees and other obstacles (Agranat, 2009; Wiley and Richards, 1978). Richards and Wiley (1980) concluded that reverberation has a more severe effect than amplitude fluctuations and its effect is lower within the 2–8 kHz range. The presence/absence of foliage and the directionality of the sound source had an influence on the amount of reverberation. They emphasized that the reverberations are severe outside 2–7 kHz and frequencies above 8 kHz are not suitable for long distance communication.

Considering these effects, the acoustic adaption hypothesis (Rothstein and Fleischer, 1987; **?**) suggests that rapid amplitude modulations (high frequency trills) and low frequency amplitude modulations (whistles) are more appropriate for open and closed habitats respectively (Brown and Handford, 2000). As we compare our findings with others, its worth mentioning that most of the other researchers located the source and

the receiver at the same height (Richards and Wiley, 1980; Marten and Marler, 1977; Ingård, 1953; Maciej et al., 2011), but we used a fixed height for the recorder (1.5m) while changing the transmission height.

Our findings of sound attenuation in relation to distance and frequency are in partial agreement with (Richards and Wiley, 1980) and (Wiley and Richards, 1978). We observed that in the open field at moderate distance, low frequencies suffered more attenuation. In contrast, in the forest, higher frequencies were attenuated more while the lower frequencies travelled further. It is clear from Fig. 3.17 that the first harmonic (just below 2 kHz) was attenuated more in the open site than in the forest. In the open field, even the furthest recorder (120m) captured 3–4 harmonics, but in the forest the furthest recorder only captured 2 harmonics. However, as demonstrated in Fig. 3.17 (a)–(b) and Fig. 3.12 we emphasise that the reverberation has lower effect within approximately 1–8 kHz rather than 2–8 kHz as approximated by Richards and Wiley (1980). The upper bound also depended on the type of bird call; for example, in saddleback, tui, and kākā songs, frequencies above 6 kHz were largely attenuated while in robin and hihi songs attenuation occurred at frequencies around 8 kHz.

The acoustic adaption hypothesis and similar suggestions of Morton (1975) were largely true in the case of ground birds. Morton (1975) suggested that narrow frequency tone-like sounds are more suitable for forest birds living close to the ground, which is true particularly for male kiwi and weka; when these calls were captured at relatively large distances (>100m) only the fundamental frequency component, and the first harmonic (<2.5 kHz) remained (see Fig. 3.17).

Ken et al. (1977) reported that habitat had less effect on sound attenuation when compared to the effect of transmission height and frequency. However our results, using real bird sounds at two heights and two differing sites, suggest that the habitat has a larger effect on birdsong capture than does transmission height (Table 3.5, Fig. 3.5, and Fig. 3.7). The selected bird sounds were less attenuated in the forest than in the open field; the only exception was the very low frequency booms that transmitted almost equally in both habitats. In the open field, even very low wind speeds can create large fluctuations and degrade the sounds at a moderate distance (Wiley and Richards, 1978). The closed forest habitat showed minimum effect of wind, yielding higher SnNR (e.g. Fig. 3.13 (b)). In addition, the presence of two parallel reflecting or refraction layers, such as the ground and canopy in the forest habitat, support the propagation of sounds generated between those two layers, sometimes making a 'negative excess attenuation', meaning that the real attenuation becomes lower than the expected attenuation (Wiley and Richards, 1978). Persistently higher SnNR in our forest experiments confirmed the advantage of the stratified media.

Atmospheric sound transmission is also affected by humidity; when the humidity

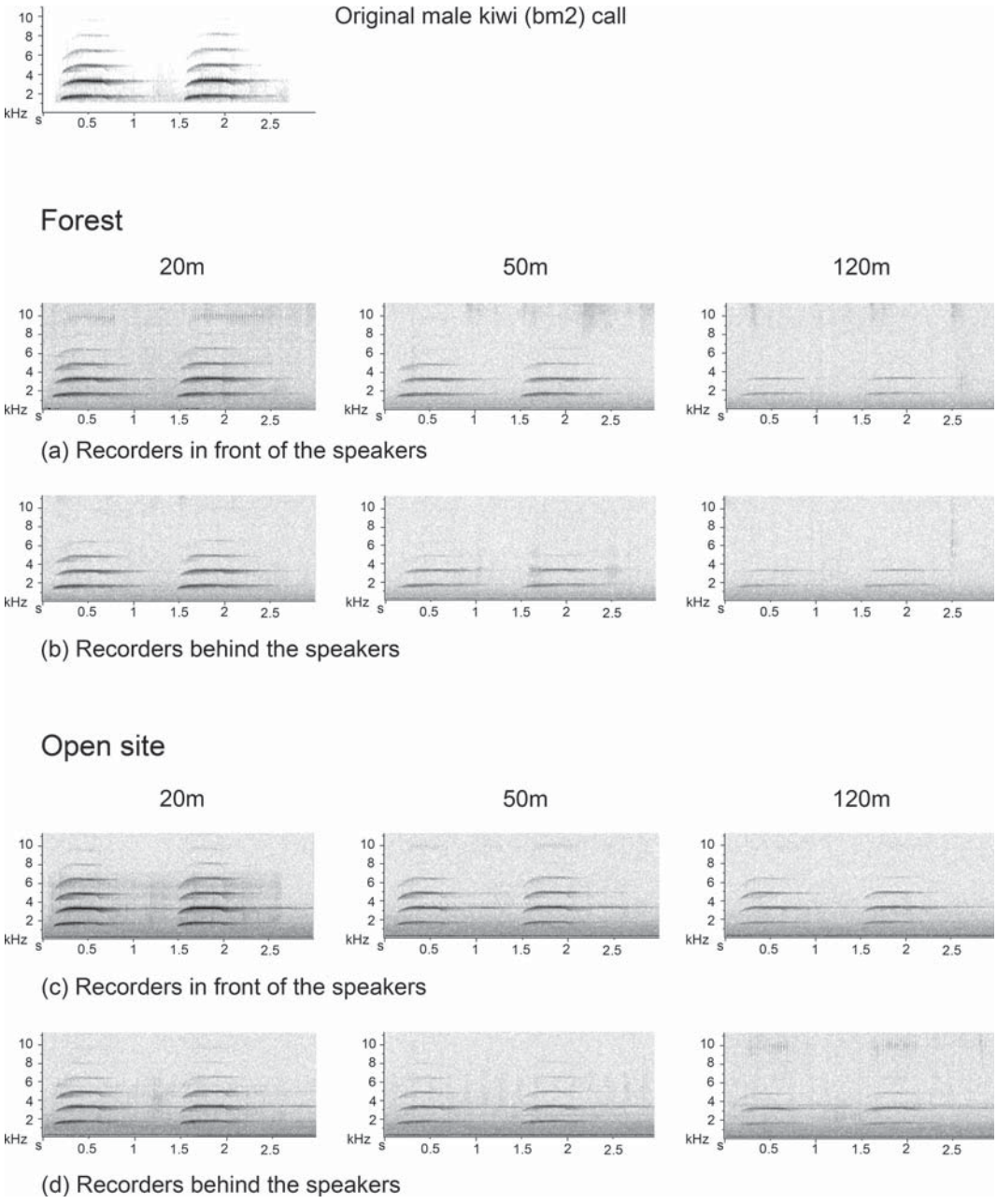


Figure 3.17: Re-captured male kiwi call (bm2) that was transmitted close to the ground in the forest and the open field at night (FNC and ONC respectively).

is lower the molecular absorption is higher and attenuation decreases approximately linearly with increasing humidity (Griffin, 1971). Even though we did not include humidity and temperature in our main analysis we measured them consistently. Between the two habitats (using four calm trials used in Analysis I), the average temperature was equal (approximately 15 °C), but the average relative humidity was higher in the open field (around 82%) than the forest (around 73%). However, as opposed to this theory, our results demonstrated that the effect of humidity was negligible compared to the effect of the wind and stratified media.

3.4.2 Sound Attenuation and the Transmission Height

Previous studies have shown that ground attenuation is highest at lower frequencies (<1 kHz), in the open space, and also depends on the transmission height and the distance to the recorder in relation to the wavelength (Ingård, 1953; Aylor, 1972; Marten and Marler, 1977; Linskens et al., 1976; Ken et al., 1977). Broadcasting white noise and pure tones (0.3–11 kHz) in temperate and tropical habitats, Marten and Marler (1977) found that transmission height had the greatest effect on sound attenuation followed by frequency at the source: there was a higher attenuation for higher frequencies. Therefore, in most cases, the lower the sound frequency the better for transmission. However Marten and Marler (1977) also found that when the sounds were produced close to the ground attenuation of very low frequencies increased. Height effect was greatest at ground level decreasing rapidly up to 2m, and levelling off with higher transmission heights. Compared to their other sites, attenuation below 2m was more significant in the open field. The variation was minimum in the forest (with leaves).

Marten and Marler (1977) suggested that this greatest attenuation at ground level in the open field is due to the interaction with the ground surface and micrometeorological events at the air-ground interface such as air turbulence and temperature gradients (these effects are higher in open habitat than forest). Further, Marten and Marler (1977) observed higher attenuation in the open field close to the ground in contrast to the less attenuation in deciduous forest with leaves and without leaves. Interestingly, at 10m height, the forest had the highest attenuation compared to the open field. They reported that at 1m transmission height, sound carries equally well in all habitats. Marten noticed a frequency range (1–3 KHz) that has minimum attenuation in all habitats at ground level (which means the window exists when the sound source is less than about 1m height from the ground). The presence of such a frequency window was first mentioned by (Morton, 1975), who conducted fairly similar experiments in the Panamanian rain forest. Somewhat different to the findings of Marten and Marler (1977), Maciej et al. (2011) experienced a stronger attenuation of primate vocalisation in the forest compared to the open habitat. They mentioned that the signal attenuation

was strongest in dense forest with low transmission height (0.5m) and lowest in the open field with high transmission height (2m).

In this study, we found higher attenuation when the speaker was close to the ground, not only in the open site, but also in the forest. Female kiwis, weka, and morepork sounds were more attenuated when the bird was close to the ground (0.25m) than at 3m height (Fig. 3.7), the attenuation in the forest was always lower than in the open field (Fig. 3.8), and the difference was higher during the night (Fig. 3.10). The calling posture of kiwi could be their natural adjustment to this: both brown kiwi (video footage of male and female kiwi) and little spotted kiwi (Digby, 2013) adopt a unique calling posture, extending the neck and pointing the bill upwards so that their sounds can probably avoid some ground attenuation. The robin is a ground forager and hihi use the ground for foraging and copulating, hence both species mainly call under the canopy. Our experiments demonstrated that robin and hihi sounds were better transmitted close to the ground, in line with their behaviour. While moreporks naturally roost on a reasonably high tree branch, this phenomena may not have major consequences to their communication. As our recorders were placed at eye level, in between these two transmission heights, we can assume that the current practice of recording of morepork is adequate. In future work we will investigate whether the sound acquisition is better at higher receiver height.

The interaction between the transmission height and the site was also significant for some other bird sounds: kākāpō chinging, hihi, and robin (Fig. 3.8). In contrast to Marten and Marler (1977), our experiments have shown that in the forest the impact of transmission height was higher than in the open site. The chinging sound of flightless kākāpō was best transmitted at ground level in the open site while transmitted equally at two transmission heights in the forest. However, it is worth noting that even though kākāpō cannot fly they are great climbers.

Marten and Marler (1977) observed that the lower the frequencies the better the transmission in any habitat (ground level to 10m source height), except that frequencies below 2 kHz were largely attenuated close to the ground. However, they broadcasted tones greater than 350 Hz because lower frequencies (<400 Hz) were not transmitted properly through their speaker system. Ground attenuation is maximum for wavelengths $\lambda, 0.1 \times h < \lambda < 0.7 \times h$, where h is the transmission height (Wiley and Richards, 1978; Ingård, 1953). In the case of low frequency kākāpō booms (centre frequency approximately 150 Hz) and bittern booms (centre frequency approximately 100 Hz) wavelength (approximately 3.4m and 2.3m respectively) did not fall into this range, perhaps this explains why the transmission height was not a factor for these two sound examples.

3.4.3 Sound Attenuation and the Directionality of the Bird and the Recorder

Richards and Wiley (1980) suggested that in scattering environments (e.g. forest) optimal directionality of sound production and reception is advantageous. But none of the researchers investigated the impact of directionality of the speaker relevant to the recorder. In this study, we found that there was a significant effect of the direction of the bird call relevant to the recorder, on birdsong acquisition. Interestingly, the recorders behind the bird (Dir3) recorded the bird sounds better than the recorders on the two other directions. We attribute this effect to the light wind, blowing from Dir1.

3.5 Conclusions

Automated recognition of birdsong is challenging in the presence of faint calls, which are a major cause of false positives (Cragg et al., 2015; Digby et al., 2013; Potamitis et al., 2014). The purpose of this manuscript was to explore how different bird calls attenuate under different environmental conditions, primarily to facilitate the development of protocols for birdsong acquisition using automated recorders. The experiments confirmed that the forest consistently caused lower attenuation compared to the open site. This result encourages the use of autonomous recorders in the forest. Song acquisition of nocturnal birds turned out to be best during the night, in line with natural selection of species song. However, the transmission height was an issue for some flightless birds resulting in more attenuation when the speaker was close to the ground, particularly for female kiwis and weka. Nevertheless, we assume that birds compensate for this disadvantage by adjusting their calling posture. Morepork vocalisations were better captured when the bird was above the ground under this recorder positioning (at eye level), confirming that the current practice of morepork data collection is suitable.

The effect of wind was severe in the open site and was minimal under the canopy in the forest. Therefore, the field recordings collected in forested habitats are less susceptible to the direct effect from the wind, but indirect affects such as boosting the rustling noise of leaves can still effect the recordings. According to the results of this study, the main advantage in the forested habitats is that bird songs are not guided by the wind, which means the recorder positioning is essentially comfortable. However, the directionality of bird calls still play a role in field recordings.

The study confirmed that higher frequencies attenuate more with distance, especially in the forest. Frequencies above 8 kHz were largely attenuated even at a moderate distance. Due to this frequency selective attenuation of bird sounds, selecting a large sampling frequency, particularly when using autonomous recorders to capture birdsong, may increase the volume of data unnecessarily. For example, when recording kiwi and

morepork we suggest that 16 kHz is more suitable and we still obtain almost the same data as at 48 kHz. That is also a way to easily avoid the species that are beyond the frequency of the target species.

3.6 Supporting Information

S 3.1 Data sheets. Analysis I data sheet

S 3.2 Data sheets. Analysis II data sheet

S 3.3 Data sheets. Analysis III data sheet

3.7 Acknowledgments

We would like to thank all the volunteers, Natasha, Tim, Sumudu, Catherine, Kim, Julia, Ross Bell, Myung Jong, Anindya, Asif, Tobby, Emma, Giulian, Janna, Kelly, and Shari, who turned up with a short notice and helped in the field. Thank you Sara Treadgold (DOC, Whanganui branch) and Emma Williams for lending enough acoustic recorders. We appreciate the great technical support given by Paul Barrett and Cleland Wallace (Technical staff, Ecology Group) particularly for making the essential setups in-house. A big thank to Gary Mack (Massey Grounds Manager) not only for facilitating to use the rugby field for playback experiments but also to kind efforts taken to minimise the machine noise by pausing the maintenance work during the trials. James Lambie (Science Coordinator, Horizons Regional Council) provided us access to the Pohangina Reserve and all the details about the tracks and safety matters in the forest. Some of the bird sounds used to playback were from Andrew Digby, Alex Brighten, Emma Williams, and Les McPherson (Natural History Unit Sound Archive, www.archivebirdsnz.com).

Chapter 4

Birdsong Denoising Using Wavelets

Abstract

Automatic recording of birdsong is becoming the preferred way to monitor and quantify bird populations worldwide. Programmable recorders allow recordings to be obtained at all times of day and year for extended periods of time. Consequently, there is a critical need for robust automated birdsong recognition. One prominent obstacle to achieving this is low signal to noise ratio in unattended recordings. Field recordings are often very noisy: birdsong is only one component in a recording, which also includes noise from the environment (such as wind and rain), other animals (including insects), and human-related activities, as well as noise from the recorder itself. We describe a method of denoising using a combination of the wavelet packet decomposition and band-pass or low-pass filtering, and present experiments that demonstrate an order of magnitude improvement in noise reduction over natural noisy bird recordings.

4.1 Introduction

More than 13% (1,373) of bird species are vulnerable or in danger of extinction from causes such as deforestation, introduction of alien species, and global climate change (International Union for the Conservation of Nature Red Data List, 2014). In order to conserve bird populations, wildlife managers require accurate information about species presence and population estimates derived from monitoring programmes. Although birds are hard to spot visually even when the observers are in the correct place, they are more vocal than other terrestrial vertebrates and therefore birdsong is usually the most direct way for humans to detect them. With the development of acoustic recorders that can be left in the field for extensive periods of time capturing all songs,

including rare ones, traditional call count surveys are being replaced by the collection of terabytes of data, which can be collected relatively cheaply and easily with limited human involvement.

The permanent storage of this acoustic data brings the advantage of being able to listen to the songs and to view their spectrograms again and again, improving the accuracy of both species recognition and call counting. However, this work is still largely manual, requiring spectrogram reading and listening, which makes it a costly approach that requires well-trained individuals; it reportedly takes an expert approximately one hour to scan the spectrogram of ten hours of recording (Digby et al., 2013) (depending on the quality of the recordings, species being monitored, and call rate), which is a daunting task, especially when many recordings are often collected simultaneously (Potamitis et al., 2014). Consequently, sampling (analysis is done on limited time periods within a subset of recordings) is favoured in many surveys, but it introduces bias and incompleteness, hence the desire to automate the recognition of bird species from their songs.

Compared to human speech recognition, one of the principal challenges of birdsong recognition—which obviously occurs in natural environments—is noisy recordings. The recorder picks up all of the noise that is in the environment, not just the birdsong, and the birds are rarely very close to the microphone. This leads to a low signal-to-noise ratio, making it hard to even detect the birdsong, let alone recognise it, whether the recognition is done by human or computer. In this paper we consider the problem of denoising birdsong from a signal processing point of view. We discuss what makes up the various types of sound that birds emit, and then the sources of noise that can be present. We then consider the signal processing methods that are available, and compare two methods: a traditional approach based on band-pass/low-pass filtering, and our own, which uses the wavelet packet decomposition in concert with band-pass or low-pass filtering. Using songs and calls from different bird species (that cover a range of vocalisations and frequencies), we demonstrate that we can significantly improve the quality of recorded birdsong, both individually segmented, and over relatively long periods.

4.2 Bird Vocalisation, Categorisation and Spectrogram Patterns

Bird vocalisations play a major role in species-specific communication, including mate attraction, parent-offspring interaction, cohesion among flocks, and territorial defence (Kroodsma et al., 1982). Experiments have shown that birds are capable of recognising conspecifics, individuals, and other species using songs alone (Catchpole and Slater,

2008). Each bird species has their own song repertoire, which can vary from monotonous repetition to innovating new, complex songs (for example, the superb lyrebird (*Menura novaehollandiae*) and brown thrasher (*Toxostoma rufum*) (Kroodsma, 2005)).

Vocalisations can be categorized into calls and songs, where calls are composed of fairly simple sounds produced by both sexes, while songs are long and complex and produced more commonly by male songbirds (order *passeriformes*). The main difference between calls and songs is arguably their function: songs are generally viewed as having a role in reproduction, while calls have an ever increasing number of functions from territoriality, to individual identification, to communicating complex messages such as type and size of predator presence (Catchpole, 1983; Morse, 1970; Lein, 1972). Songs and calls can be further divided into phrases, syllables, and elements (Somervuo et al., 2006), as shown in Fig. 2.2 (a). The fundamental unit of sound is the element, with syllables being comprised of one or more elements that can be separated from the other content of the vocalisation. A series of syllables that are organised into some pattern is referred as a phrase.

There are a number of studies that define the components of bird vocalisations based on the patterns they generate in a spectrogram (Duan et al., 2013, 2011). The key acoustic components defined by Duan et al. (2013) are lines (at any angle), warbles, blocks, oscillations and stacked harmonics (examples are given Fig. 2.2 (b)-(k)).

4.2.1 Spectrogram Analysis

The spectrogram representations of birdsong shown in Fig. 2.2 are based on the frequency representation of a discrete recording of the continuous birdsong. Digital recording of birdsong is based on equally-spaced time sampling of the analogue birdsong. This primary form of acoustic data are referred to as the oscillogram or simply the waveform. The oscillogram is two dimensional: the horizontal axis represents time and the vertical axis represents amplitude. It turns out that signal analysis is generally more effective in the frequency domain than in the time domain, as is evidenced by the fact that ornithologists prefer the spectrogram representation to the oscillogram one. The frequency representation provides information about the frequency components that comprise the signal, but not about when those frequencies occur. Converting the waveform into the frequency domain is performed by the Fourier transform, which represents the signal as a weighted combination of sine and cosine waves at different frequencies. The Fourier transform is invertible, meaning that processing can be performed in the frequency domain and then transformed back into the time domain, for example to enable the sound to be played.

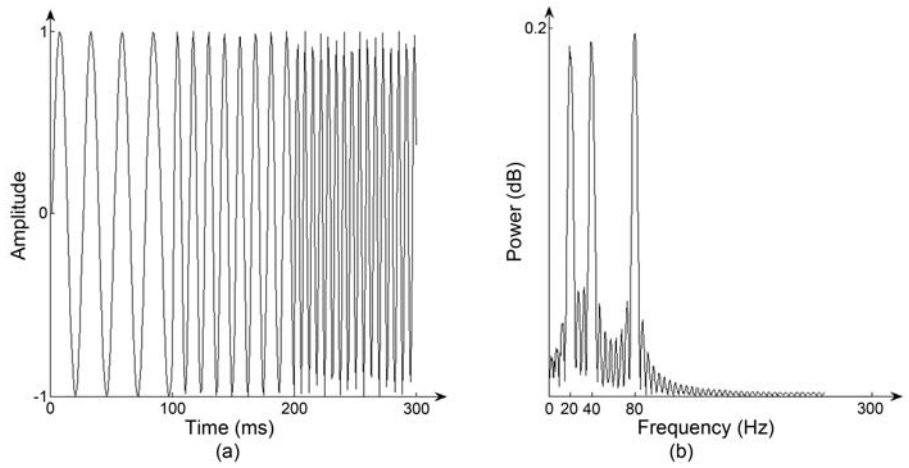


Figure 4.1: Non-stationarity. (a) A non-stationary signal containing 20 Hz, 40 Hz and 80 Hz frequencies and (b) its power spectrum computed using the Discrete Fourier Transform.

Birdsong is transferred into the frequency domain by applying the Discrete Fourier Transform (DFT), and in practice the Fast Fourier Transform (FFT), which is a computationally efficient algorithm for the DFT, is used.

Fig. 4.1 (a) shows a non-stationary signal. During the first 100 ms the frequency of the signal is 20 Hz, during the second 100 ms the frequency doubles and again during the last 100ms. The right of the figure shows the *power spectrum*, which plots the energy per time unit (power) against the frequency components, and which clearly shows the basic frequencies of the original signal. Thus the power spectrum is a good representation of sound, summarising its periodic structure. However, it is suitable only for stationary signals while most signals in the real world are transient (non-stationary). The reason for this is that the signal is assumed to be infinite in time, and choosing a short time window has the effect of causing aliasing, where signal from outside the chosen range affects the appearance inside the range.

Segmenting the entire signal into fixed size small time windows and then calculating frequency components from these windows is a common practice based on the assumption that the signal is stationary over a short duration. Careful use of windows that decay to zero at the edges of their range and overlapping the windows enables Short Time Fourier Transformation (STFT) to be used, and this is the basis of the spectrogram. First, the power spectrum of each window is calculated, and then rotated 90°, and the amplitude is replaced by a greyscale. The complete spectrogram is generated by stacking all those images of subsequent windows appropriately. Provided that the time windows are short enough that the frequency components are stable in the time

window this provides a faithful representation of the frequency components of the data against time, but it comes at a cost, since estimating frequencies accurately requires time: frequency resolution can only be achieved at the cost of time resolution and vice versa. The result of this is that larger windows are required for low frequencies, but STFT cannot deal with these subtleties. This led us to consider wavelets as a representation of birdsong, as we shall discuss after we consider the types of noise that are present in birdsong recordings.

4.3 Bird Recording and Noise

Until recently, manual (attended) recording was the method of choice for recording birdsong. This generally enables the capture of good quality close-range songs provided the recordist has the skills not only to tune and handle the recorder, but also a good knowledge of the bird being recorded and how to approach it closely. The advent of waterproof programmable recorders with good battery life and high recording capacity has enabled a new form of birdsong recording, enabling ecologists to collect every sound in the forest (or other area of interest) without disturbing the birds or requiring groups of experts to perform call counting in the field. However, recordings made in natural environments are highly susceptible to a variety of noises. During attended recordings, some noise can be controlled by careful screening, but in automatic recording this is impossible.

4.3.1 Types of Noise

The sounds that can be heard can be categorised into three broad types: biophony, geophony, and anthrophony (Farina, 2014). Biophony refers to any sound produced by biological agents: in the forest major biophonies are birds, insects, frogs/toads, and mammals. Because we are only interested in acoustic activity of birds, all other biological sounds are categorised as noise; with recordings targeted at particular bird species, even other birdsong is regarded as noise. Geophony refers to all non-biological, natural sounds in the environment such as wind and its effect on trees, rain, thunder, and running water. Field recordings are always blended with these geophonies. Anthrophony refers to all sound generated from human-made machines such as aircraft, vehicles, wind turbines, and the recording device itself: there is always some microphone and recorder hum. Collectively, these noises contaminate all acoustic data to a greater or lesser extent, see Fig. 4.2 (a)-(b). The problems of noise are both that it can mask the signal of the bird call, and also transform it so that it looks different, making it hard to identify. While there is some research on features that are invariant to noise, meaning that they look the same even in noisy data, they are not general, and we will

not consider them further here.

We differentiate between denoising of a signal, which is principally the removal/filtering of consistent noise, from source separation, which is identifying that there are several birds calling simultaneously and separating the signals into individual birds. We do not consider the second further in this paper; (Pedersen et al., 2007) provides a survey of approaches to the problem, but notes that very few of the methods have been shown to work for real-world signals.

There is a theory of noise in digital signal processing (see, for example Vaseghi (2008)), which characterises the noise according to its properties into:

White noise has equal energy at all frequencies, meaning that the power spectrum is flat. In practice, noise is only white over a limited range of frequencies (Fig. 4.2 (e)). While not all white noise is Gaussian, natural white noise can often be modelled as such.

Coloured noise shows a non-uniform power spectrum, with the energy generally decreasing in proportion to the frequency f. Common types of coloured noise include pink (power $\propto \frac{1}{f}$) and brown (power $\propto \frac{1}{f^2}$).

Impulsive noise refers to sudden click like sounds that last for a very short period of time (milliseconds), such as switching noise. An ideal impulse generates a horizontal line in the power spectrum because these sharp pulses contain all frequencies equally.

Narrow-band noise such as microphone hum shows a small range of frequencies.

Transient noise is a burst of noise that occurs for some time, and then disappears.

An important property of any sound is whether or not it is stationary i.e., its properties do not change substantially over time. Most noise in natural recordings is at least quasi-stationary, being geophonic in nature. However, birdsong is not stationary (i.e., it is transient) since it is generally short-lived and varies quickly. This difference between the properties of the noise and signal enables noise reduction techniques to be applied.

4.3.2 Noise Filtering

Noise filtering is the most common approach to dealing with noisy recordings. Traditional signal processing, based on electronics, uses two basic filters, low-pass and high-pass, which allow frequencies respectively below and above a pre-defined cut-off frequency to pass through, and attenuate the rest. Combining a low-pass filter and a high-pass filter gives a band-pass filter. If the noise occupies high frequencies while

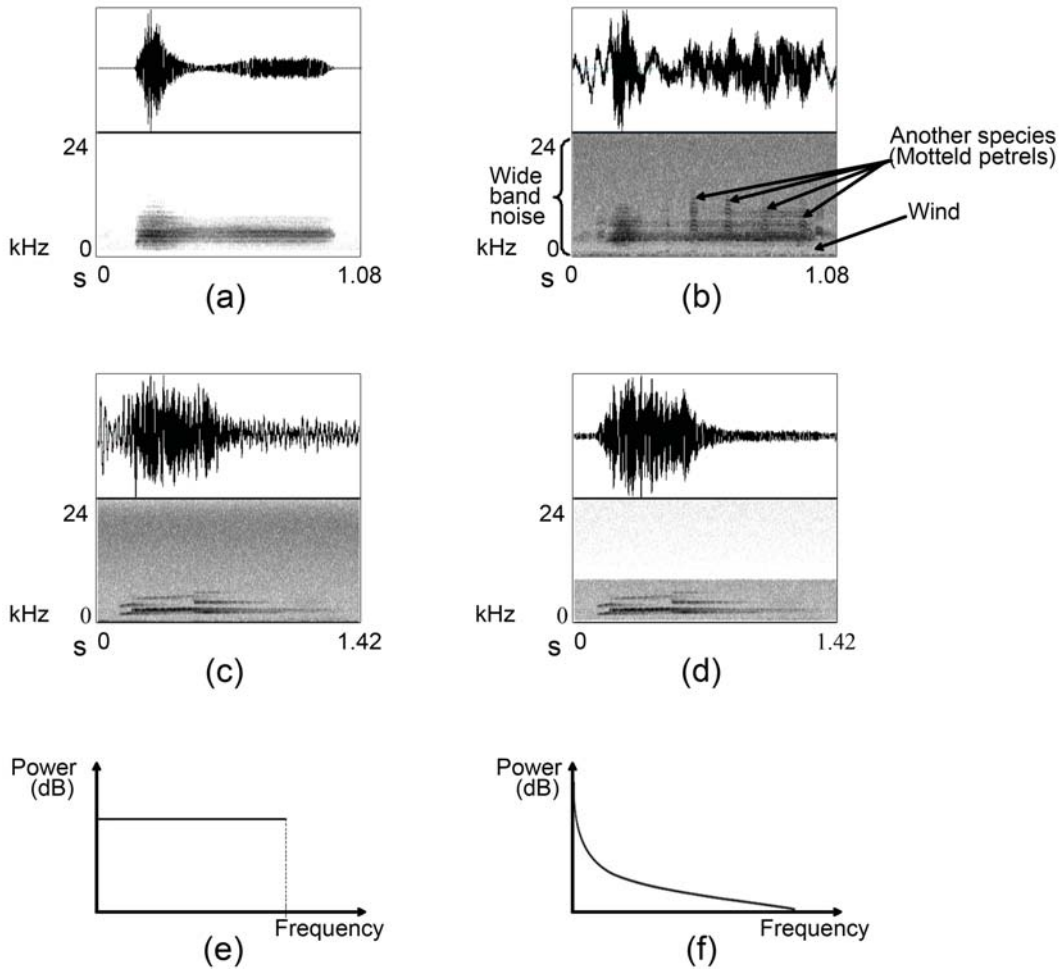


Figure 4.2: Examples of bird calls with various degrees of noise, the effect of band-pass filtering and power spectrum of white and pink noise. The top row of each sound figure displays the oscillogram and the second row the spectrogram. (a) A less noisy example of kākāpō *chinging* with limited noise and (b) a noisy example of kākāpō *chinging*. (c) An original male kiwi whistle and (d) its noise filtered (band-pass) signal. Noise is visible as a grey background in the spectrogram surrounding the sound depiction and most of the high-frequency variation in the oscillogram. Power spectrum of (e) white noise and (f) pink noise.

the bird of interest sings low frequency songs then this would be sufficient to eliminate noise, but since the spectra of the noise and the signal overlap, this is not the case.

Fig. 4.2 (c)-(d) illustrates the effect of band-pass filtering on a single instance of a male North Island brown kiwi (*Apteryx mantelli*) whistle. The spectrogram shows that all the high frequency and low frequency noise components have been removed successfully, but all the noise in the range of the bird's song frequency (visible as grey background) is still there, confirming that this basic filtering is not sufficient to recover birdsong. Further, birds have different call categories from different frequency bands. For example, the kākāpō (*Strigops habroptilus*) generate two types of vocalisation: *booming*, which is a very low frequency call and *chinging*, which is a relatively high frequency call. Designing a common filter to clean hours of kākāpō recordings is impossible because they do not share the same frequency range.

Another traditional approach is the Wiener filter, which generates an estimate of the desired or target random (Gaussian) process based on linear time-invariant filtering and the minimum mean square error between the estimated signal and the desired signal by assuming that the signal and noise are stationary and spectral information is available (Vaseghi, 2008). This is not true for birdsong, therefore we did not consider it further here.

4.4 Wavelets

We explained earlier that the Fourier transform, while commonly used in birdsong analysis, is not really suitable because of the tradeoff between temporal resolution and frequency resolution. An alternative is the *wavelet transform*, which is a relatively recent development in signal processing (Morlet et al., 1982), although it has been invented independently in fields as diverse as mathematics, quantum analysis and in electrical engineering (Mertins, 2001). Wavelets have been applied in many areas, such as data compression, feature detection and denoising signals (Graps, 1995).

In the Fourier transform the signal is mapped into a basis of sine and cosine waves. The wavelet transform also uses a basis, but the basis elements are scale-invariant, meaning that they look the same at all scales, and they are localised in space. The upshot is that in the wavelet representation different window sizes can be used to see the signal at different resolutions; an analogy would be viewing a forest and its trees at the same time. If we need to see the whole forest we have to see it at a large scale and then we can capture global features. In order to see the trees, we have to zoom in and to focus on a tree. Zooming more allows us to see leaves. We can see the forest, trees and even leaves by using different scales. Fig. 4.3 (a)-(b) highlights the difference between Fourier and wavelet analysis: the window size in Fig. 4.3 (b) is more flexible (allowing large windows for low frequencies and small windows for higher frequencies),

which is important for broad spectrum non-stationary signals such as birdsong.

There are several choices of basis features (referred to as *mother wavelets*) Ψ, and unfortunately the best mother wavelet for a particular application needs to be determined experimentally. Fig. 4.3 (c)-(e) shows examples of some mother wavelets, including the simplest Haar wavelet, which is a discontinuous step function. While the discontinuity can be a disadvantage in some domains, including birdsong, it is beneficial for those that exhibit sudden transitions like machine failure (Patil et al., 2014). Fig. 4.3 (d) provides three examples of the Daubechies wavelets (dbN) showing that the smoothness of the wavelets increases as N increases. Finally Fig. 4.3 (e) shows the discrete Meyer wavelet ($dmey$).

In order to construct other elements of the wavelet basis the mother wavelet is scaled and translated by factors a and b using:

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\Psi\left(\frac{t-b}{a}\right) \tag{4.1}$$

Parameter $a \neq 0$ determines the amount of stretching or compression of the mother wavelet (depending whether a is greater than or less than 1). Therefore, when a is small high frequency components are introduced to the wavelet family; in return those wavelets can capture high frequencies of the signals. In the same manner, when a is large low frequency components are introduced to the wavelet family and help to capture low frequency signals. Parameter b determines the amount of shifting of the wavelet along the horizontal axis: $b > 1$ shifts the wavelet to the right, while $b < 1$ shifts it to the left. Therefore, parameter b specifies the onset of that wavelet. Fig. 4.3 (f) illustrates the effect of a and b with respect to a given mother wavelet. Accordingly, wavelets are defined by the wavelet function (*mother wavelet*) and scaling function (also called the *father wavelet*). The scaled wavelets are known as *daughter wavelets*.

4.4.1 Wavelet Packet Decomposition

When wavelets are applied to a discrete signal, low-pass and high-pass filters are used, splitting the data into a low frequency (approximation) part and a high frequency (detail) part. These filtered representations of the data can then be analysed again by a wavelet with smaller scale by creating a new daughter wavelet, typically at half the scale. One modelling choice that can be made is whether to reanalyse both the approximation and detail parts of the signal, or just the approximation coefficients. We choose to analyse both, in what is known as the wavelet packet decomposition (Burrus et al., 1997). It leads to a tree of wavelet decompositions, as shown in Fig. 4.3 (g), and provides a rich spectral analysis, since there are 2^N leaves at the base of the tree when there are N levels.

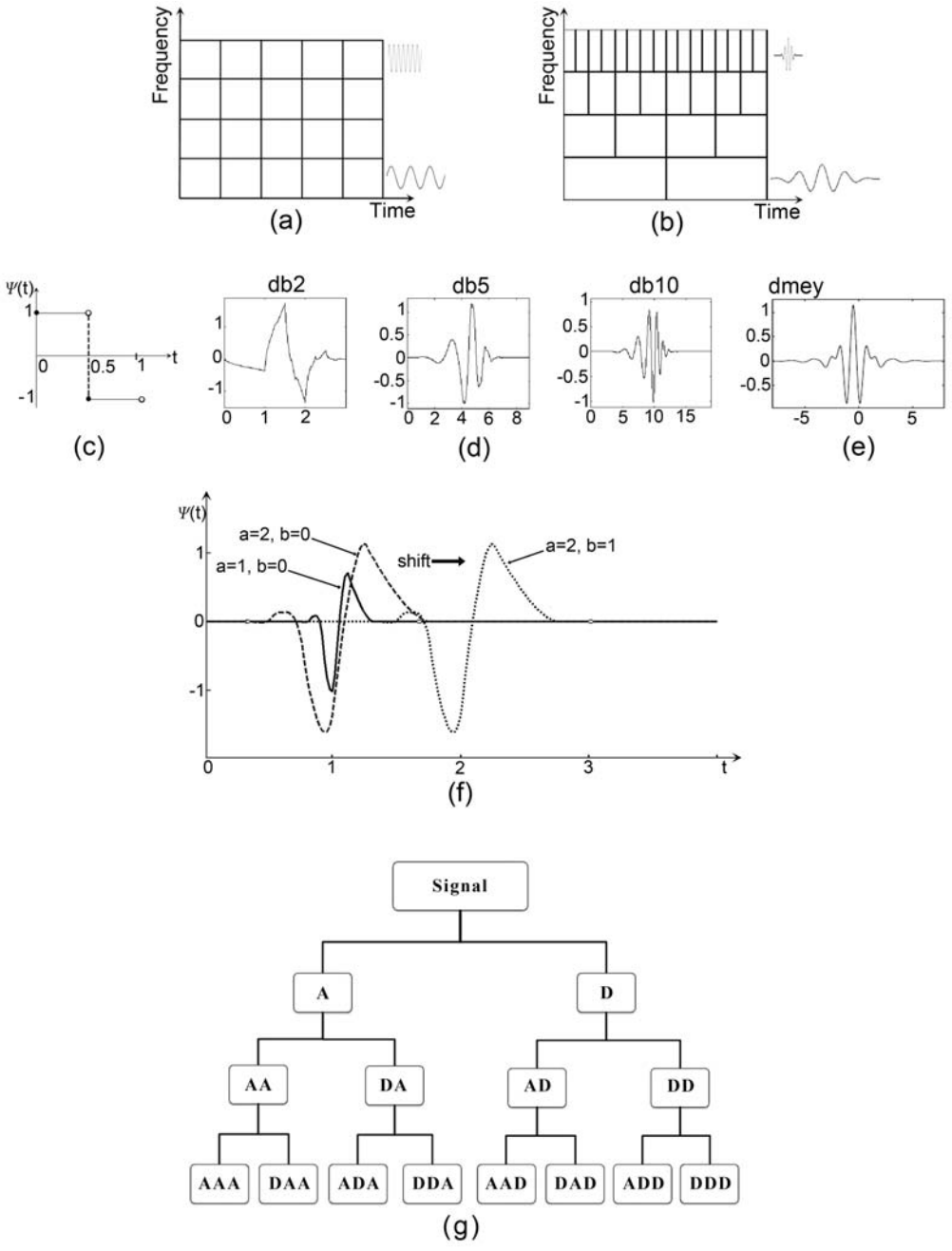


Figure 4.3: Wavelets and their relation to time-frequency resolution and wavelet packet decomposition. Time-frequency resolution in (a) STFT and (b) wavelets. Examples of mother wavelets: (c) *Haar*; (d) a subset of Daubechies wavelets; (e) Discrete Meyer wavelet. (f) Scaling and shifting the mother wavelet $\Psi_{1,0}(t)$ gives two new wavelets $\Psi_{2,0}(t)$ and $\Psi_{2,1}(t)$. (g) A level three wavelet packet decomposition tree (A=approximation and D=detail).

However, the question of how many levels to use in the tree still remains. This question is often answered experimentally, but since we want a method that can work unaided on birdsong, we need to find a computational approach. We have approached this by considering how much information about the signal is contained in the approximation at each node, reasoning that nodes that do not contain information are representing the noise, and so should be discarded. In the field of information theory, Shannon entropy provides the standard measure of uncertainty or disorder in a system (Shannon, 1948), and this is connected to the amount of information contained in a given signal (Marsland, 2014).

The entropy S of a set of probabilities p_i is calculated as (using the convention that $0 \log 0 = 0$):

$$H(p) = -\sum_i p_i \log_2 p_i \tag{4.2}$$

where p_i is the probability of i^{th} state in the state space. In wavelets, we used a slightly different version of this Shannon entropy:

$$S = -\sum_i s_i^2 \ln(s_i^2) \tag{4.3}$$

where s_i is i^{th} sample of the signal (Wang et al., 2011; Ma et al., 2012).

The idea of using entropy for wavelets is to argue that when the entropy is small, the accuracy of the selected wavelet basis is higher (Ma et al., 2012). We used this computation at each node to choose whether or not to retain a node, and stopped creating the tree at the point where all of the nodes contained noise are removed by this computation, meaning that the signal was fully described.

4.4.2 Previous Uses of Wavelets for Bioacoustic Denoising

The use of wavelets for noise reduction, referred to as denoising, is still an emerging advance in digital signal processing. While there are some examples of denoising in other audio signal domains such as partial discharges (PD) signals (Ma et al., 2002; Shim et al., 2001; Tsai, 2002), music (Sharma and Pyara, 2013), speech (Bee et al.), and phonocardiography (Vaisman et al., 2012; Varady, 2001), their use in bioacoustic denoising is still uncommon. In addition, the two studies we know of which used wavelets for denoising animal sounds did not use natural noise, but added manual noise to their recordings. Gur and Niezrecki (2007) denoised West Indian manatee (*Trichechus manatus latirostris*) vocalisations with added boat noise, while Ren et al. (2008) attempted to denoise vocalizations of the ortolan bunting (*Emberiza hortulana*), rhesus monkey (*Macaca mulatta*), and humpback whale (*Megaptera novaeanglia*), with

added white noise.

However, wavelets have been used for birdsong recognition: (Selin et al., 2007; Turunen et al., 2006) used the wavelet packet decomposition to extract features from birdsong from eight species. Interestingly, in Turunen et al. (2006) they added noise filtering via either a low pass filter or an adaptive filter bank with eight uniformly spaced frequency bands. These filtered signals were also analysed by wavelets and compared for recognition accuracy with the unfiltered version. In addition, Chou and Liu (2009) used wavelets to represent birdsong in conjunction with Mel Frequency Cepstral Coefficients (MFCCs) for recognition of 420 bird species; but the dataset in their experiment was very limited, with only one recording per species (half of each birdsong file for training and the remaining for testing).

4.4.3 Our Algorithm

To summarise our approach to birdsong denoising, we took the following steps, which are discussed further next:

1. Find a suitable mother wavelet.
2. Find the most suitable decomposition level based on the Shannon entropy.
3. Apply the wavelet transform to the noisy signal to produce the noisy wavelet coefficients.
4. Determine the appropriate threshold to best remove the noise based on the Shannon entropy.
5. Invert the wavelet transform of the retained wavelet coefficients to obtain the denoised signal.
6. Apply a suitable ordinary band-pass or low-pass filter where possible to remove any noise left outside the frequency range of the signal.

Selecting the Mother Wavelet Choosing an appropriate mother wavelet is the key to the successful estimation of the noiseless signal. One approach is to visually compare the shapes of the mother wavelets and small portions of the signal, choosing the wavelet that best matches the signal (Shim et al., 2001). However, given that we want the method to work with a wide variety of different bird calls, eyeball selection is not sufficient.

Another approach is based on the correlation between the given signal and its denoised signal (Ma et al., 2002), reasoning that if two signals are strongly linked they should have high correlation. Therefore, we can expect that the optimum wavelet

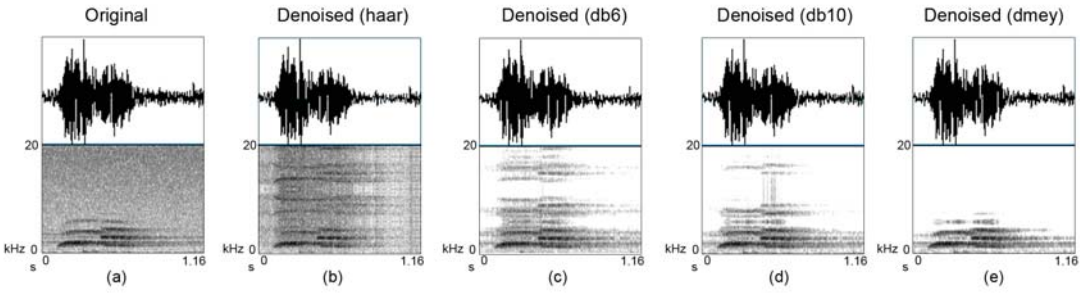


Figure 4.4: Different mother wavelets produce different results. Same excerpt of a male kiwi whistle (a) original whistle and (b)–(e) denoised with different mother wavelets.

maximises the correlation of initial signal and denoised signal. We can compare the correlation under different wavelets and pick the wavelet that generates the highest correlation. Accordingly, we analysed the correlation given by different wavelets including the Daubechies wavelets (dbN, where N represents the order) and the Discrete Meyer wavelet ($dmey$). Initial experiments showed that the $dmey$ wavelet generated the highest correlation. For instance, $db2$ (0.9950) was better than $db1$ (0.9884), $db6$ (0.9970) was better than $db2$, $db10$ (0.9971) was better than $db6$, and $dmey$ (0.9973) was better than $db10$. Then, we investigated the spectrograms of the denoised examples in order to see the actual improvement of the songs. Visual inspection (for example Fig. 4.4) also confirmed that the $dmey$ wavelet (Fig. 4.3 (e)), successfully denoised the songs without distorting them with a selection of different birdsong, and so we used that for the rest of our experiments.

Selecting the Best Decomposition Level Because we used Shannon entropy to choose the decomposition level, different birdsong will produce trees of different depths: less complex birdsong will have small trees, while more complex birdsong will require larger trees. In fact, even within single types of call, different depths of tree can be seen. We therefore ran the depth selection algorithm on every birdsong individually; while this is computationally expensive, it does lead to significantly better results. Methods to speed up this approach will be investigated in future work. So far we found that the top-down approach (start with a small tree with level 1 and expand it based on the Shannon entropy) is more efficient than the bottom-up approach (start with a big tree and shrink it); therefore we used the top-down calculation here. Starting from level 1, decomposition was continued until the maximum entropy of a parent node (at level L) was lower than the maximum entropy of its child nodes (at level $L+1$). At that point the decomposition was stopped, and the best decomposition level was determined as L.

Selecting the Threshold Each node in the decomposition tree is represented by its wavelet coefficients, and the 'impurity' of those nodes can be calculated using (Shannon) entropy. Then, eliminating noisy nodes is done by applying a threshold to each node. There are two forms of thresholding methods: *hard thresholding* and *soft thresholding*. In hard thresholding, sometimes called the 'keep or kill' method (Aboufadel and Schlicker, 2011), coefficients are removed if they are below a previously defined threshold. In contrast, soft thresholding shrinks the wavelet coefficients below the threshold rather than cutting them off sharply. Soft thresholding provides a continuous mapping and in our case it demonstrated better noise reduction without information loss yielding high SnNR (this term will be defined in the section on evaluation metrics) in initial experiments. Therefore, we used soft thresholding here.

The challenge of setting the threshold remains, however: ideally, the selected threshold should achieve satisfactory noise removal without significant information loss. If the selected threshold is too high, then it removes too many nodes from the tree, resulting in a denoised signal with missing information, while if the threshold is unnecessarily low, it does not remove all the noisy nodes, resulting in a signal that still has noise in. There will be no globally optimal threshold, and so we again selected it based on analysis of each birdsong. As was mentioned previously, many types of noise can be approximated as having a Gaussian distribution, and this is more obvious in the high frequency parts of the spectrum. We therefore computed the standard deviation of the lowest level detail coefficients in the tree, and used 4.5 standard deviations as the threshold, which should cover 99.99% of the noise (Tsai, 2002).

4.5 Experimental Evaluation

In this section we compare our wavelet-based algorithm with traditional band-pass or low-pass filtering. We introduce our dataset, and the metrics that we use to compare the results, before demonstrating the results.

4.5.1 Datasets

Primary Dataset Initially, three manually generated pure sound examples (one impulsive 'click' sound and two tonal combinations) were used to examine the performance of the proposed method against white and different coloured noises. These examples were separately polluted with different levels of these noises manually and then denoised to eliminate the noise.

Secondly, songs of two endangered and one relatively common New Zealand bird species were considered: North Island brown kiwi, kākāpō, and ruru (*Ninox novaeseelandiae*). Most of the recordings were collected using automated recorders, but a few

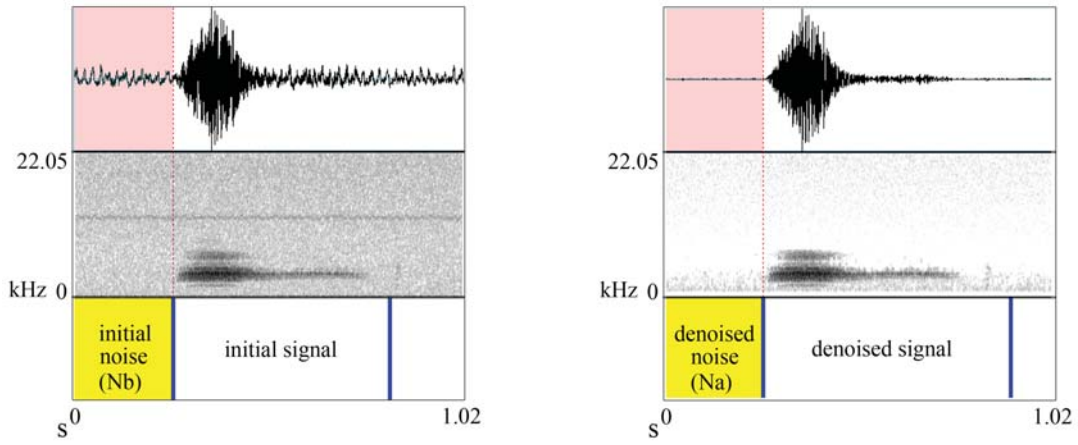


Figure 4.5: An example of kākāpō *chinging* used in the experiment. Top, middle, and last rows represent oscillogram, spectrogram, and labels indicating the parts of the recording used to calculate the SnNR respectively. (a) Initial signal and (b) the same signal after denoising and band-pass filtering.

were recorded manually. Most ruru and kiwi calls were obtained by the authors, while some ruru and all kākāpō calls came from other sources (see the Acknowledgements). The spectrogram patterns of these species are shown in Fig. 2.2. Birdsong were segmented manually into syllable level components (e.g., Fig. 4.5). The dataset (available at `http://avianz.massey.ac.nz`) contained a total of 700 syllables from seven basic call types, 100 of each (Table 4.1). These recordings were polluted with different types and levels of noise while recording. Mainly the noise was wide-band; sometimes it was concentrated more to low frequencies (for example due to wind and aeroplane noise) others to high frequencies and/or to narrow bands (for example due to insect noises like crickets and weta).

Secondary Dataset We tested our algorithm on a secondary dataset because our primary dataset did not cover all possible spectrogram patterns we expect to see in recordings collected in the wild. The songs in this dataset were mostly collected by the authors using manual recorders, but the selected recordings include significant amounts of noise. The kākā and tui songs were recorded using automated recorders by others. The eight species in this dataset (see Table 4.2) comprise seven song birds and one parrot, which have complex songs and great song diversity. We used whole songs instead of syllable level components. Five noisy song examples of each species were used, except for hihi; this species has very short songs and therefore we used ten examples.

Table 4.1: List of species, their call types and frequency range. We use the common names given by researchers for the different types of calls.

Species/call type	**Observed frequency range (Hz)**
North Island brown kiwi	
male	500–8,000
female	500–6,500
Ruru	
trill	500–8,000
more	500–2,000
pork	500–2,000
Kākāpō	
booming	0–800
chinging	1,000–12,000

Further, we were interested to see the performance of this technique over unsegmented recordings. Therefore, we denoised five series of consecutive calls from each call type from each species mentioned in the primary dataset. Then we compared the calls in the denoised series to their respective segmented calls.

Another concern when denoising birdsong is the effect of overlapping bird calls. To test this issue, we selected ten examples of recodings that contained overlapping songs from different combinations of species. Examples include overlapped male kiwi-female kiwi, male kiwi- ruru trill, male kiwi-more-pork, two of more-pork-trill, two of male kiwi-female kiwi-more-pork, tui-more-pork, robin-tui, and kākāpō chinging-mottled petrels. Again the dataset is available at `http://avianz.massey.ac.nz`.

4.5.2 Evaluation Metrics

The main measurement of true interest in denoising is the Signal-to-Noise Ratio (SNR), which can be calculated by dividing the power of the signal (S) by the power of noise (N), as given in equation 4.4, which is in units of decibels (dB). The higher the value of the SNR, the less noisy the signal.

$$SNR = 10 \log_{10} \left(\frac{S}{N} \right) \tag{4.4}$$

The challenge for real-world applications such as birdsong is that the signal and noise are not actually known because they are together in the recording. This means that computing S and N is not actually possible. Under the assumption that the noise is relatively stationary, we have estimated the power of the pure noise by isolating parts of the recording without birdsong, which should theoretically be silent, and modified

Table 4.2: List of species introduced to the secondary dataset and their song characteristics.

Common name	Scientific name	Observed frequency range (Hz)	Song structure
North Island robin	*Petroica longipes*	1,700–12,500	Males sing loud songs that have series of phrases. Phrases have variety of simple notes.
Tui	*Prosthemadera novaeseelandiae*	400–18,000	Loud and complex songs: mix of melodious notes with coughs, grunts and wheezes.
North Island kākā	*Nestor meridionalis*	700–15,000	Harsh and grating sound, variety of musical whistles.
Hihi	*Notiomystis cincta*	1,000–21,000	Variety of 2–3 note whistles. Quiet or aggressive warbles.
North Island saddleback	*Philesturnus rufusater*	800–22,000	Very active and noisy. Loud chattering calls and variety of rhythmical songs.
Marsh wren	*Cistothorus palustris*	500–15,000	Gurgling and rattling trill.
Western meadowlark	*Sturnella neglecta*	650–12,500	Male sings a complex, two-phrase song, begins with 1–6 pure whistles then a series of 1–5 gurgling warbles.
Horned Lark	*Eremophila alpestris*	1,100–18,000	Musical songs: fast, high-pitched sequence of sharp, tinkling notes.

equation 4.4:

$$SnNR = 10 \log_{10} \left(\frac{S+N}{N} \right), \tag{4.5}$$

where $S + N$ is the power in the initial signal. By comparing this computation with the denoised version we can see how effective the denoising is. Fig. 4.5 illustrates the calculation. Notice that to be able to calculate the initial noise and denoised noise we segmented the recording leaving a small period of silence at the beginning and/or end of the bird call. A comparison of original SNR and respective SnNR are shown in Fig. 4.6, where noise and signal are known.

If we recall that the noise is approximately Gaussian, a second possible metric is to measure its statistical properties, particularly its variance, reasoning that successful denoising should substantially reduce the variance of the noise. We used the same segments of 'pure' noise in the signal as were used to estimate the power of the noise

in the SnNR to compute the variance of the noise before and after denoising, terming this measure the *success ratio*:

$$\text{Success ratio} = \log_{10}\left(\frac{\text{var}(Nb)}{\text{var}(Na)}\right), \tag{4.6}$$

where Nb is the initial noise and Na is the noise after denoising. If the success ratio is greater than 0, it implies that song denoising has been successful.

A third possibility is to calculate the Peak Signal to Noise Ratio (PSNR), a widely used objective quality metric in image and video processing (Hore and Ziou, 2010; Huynh-Thu and Ghanbari, 2008). PSNR looks only at the peak value the signal can reach and the mean-squared error between the reference and the test signals. Here we used a modified PSNR (Najafipour et al., 2013) to compare noise reduced songs with their original noisy version.

$$PSNR = 10\log_{10}\left(\frac{MAX_{sig}^2}{MSE}\right) = 20\log_{10}\left(\frac{MAX_{sig}}{\sqrt{MSE}}\right), \tag{4.7}$$

where MAX_{sig} is the maximum value of the reference signal and MSE is the mean-squared error. In this calculation we maintained the noisy song as the reference and its recovered song as the test. PSNR will be relatively lower if the song is less cleaned and higher if the song is well cleaned.

4.6 Results

We implemented our algorithm in Matlab® using the Wavelet Toolbox™, which is a comprehensive toolbox for wavelet analysis. The code is available at: `http://avianz.massey.ac.nz`. As an initial experiment, white noise, pink noise, and brown noise were added to selected tonal and impulse sounds separately as a percentage of the strength of the signal. These noisy examples were cleaned using the proposed denoising approach (steps 1-5 only; without filtering), and the calculated SNR and SnNR of noisy and recovered songs are plotted in Fig. 4.6. Here we can calculate the SNR of the noisy examples perfectly because we know the actual noise added as well as the pure signal. A comparison of conventional SNR and SnNR is illustrated in Fig. 4.6 (a) confirming that both metrics perform almost equally. The same figure also shows that even in the presence of high levels of white noise, denoising using our approach is very successful. Fig. 4.6 (b) reveals that the proposed denoising approach can deal well with pink noise, but not to the extent of white noise. However, denoising brown noise still remains a challenge as shown in the Fig. 4.6 (c). This is because of its strong non-Gaussianity.

Each call example in both primary and secondary datasets was treated with three approaches: band-pass or low-pass filtering alone (F), wavelets alone (D), and wavelets

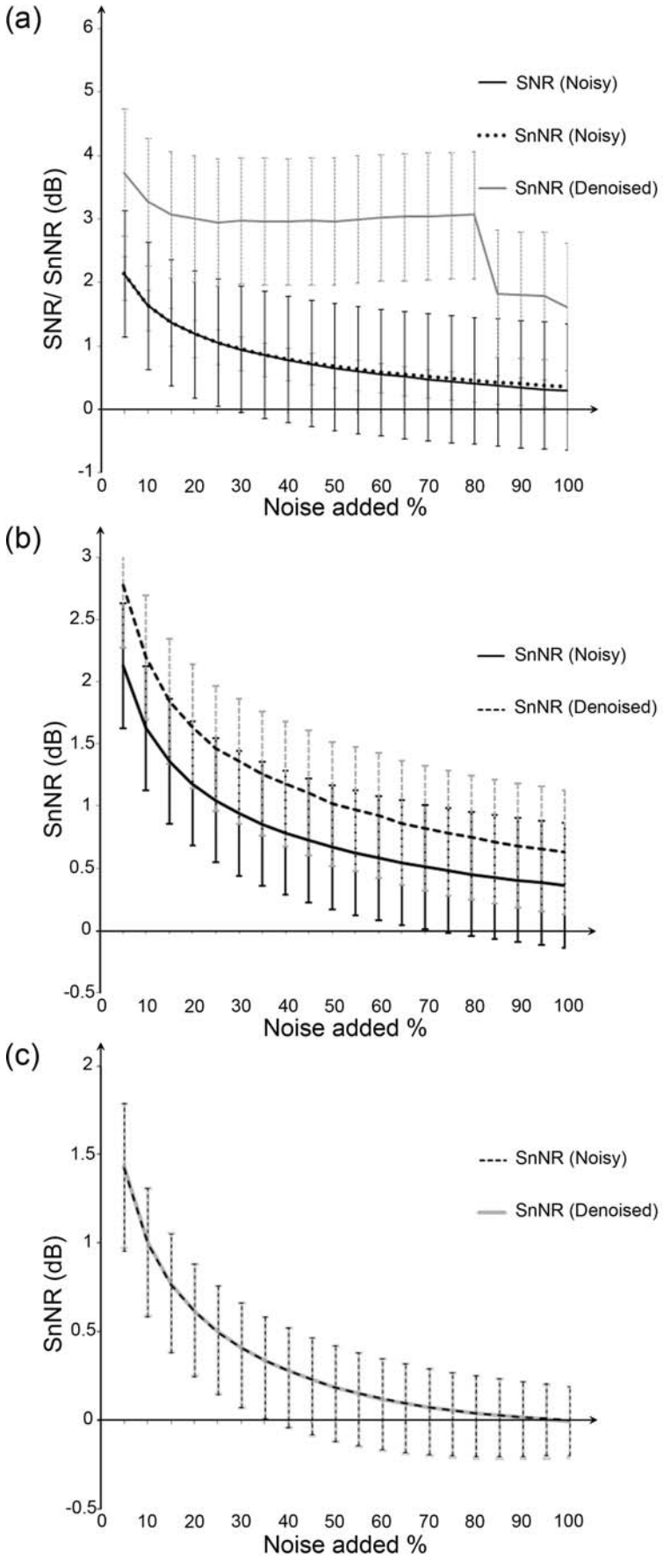


Figure 4.6: Denoising different types of noise. (a) White noise, (b) pink noise, and (c) brown noise.

and band-pass or low-pass filtering (DF). In the case of filtering, the frequency bands were selected according to Table 4.1 and Table 4.2. Fig. 4.7 (S 4.1 Audio) demonstrates that our algorithm removed most of the noise from the birdsong while preserving most of the song information. Success is visually clear from the spectrograms, for example if we consider Fig. 4.7 (a), almost all the background grey colour (caused by noise) in the original kiwi whistle has been eliminated, while the five original harmonics are still present without distortion after denoising. We examined visually and aurally each example individually to confirm whether they were improved after denoising, and found that all the calls were significantly improved. The improvement in the sound quality of the songs was successfully reflected by SnNR and Sussess Ratio (Table 4.3 and Fig. 4.8). The overall SnNR improved from 0.667 to 3.506, an improvement of more than 5 times while SnNR improved only up to 1.526 after conventional filtering. Success ratios after filtering alone and with denoising were 1.071 and 2.170 respectively. Parallel to this, PSNR increased from 10.428 to 10.694. While we have included PSNR (equation 4.7) in our results, we do not believe that it is a particularly useful measure. First, the numerator uses the maximum amplitude (hence the 'peak' in the name), which does not change when the signal is denoised, while the denominator is the root mean square of the error, which is small. This leads to a less sensitive measurement. For example, denoising alone always generated the highest PSNR because the oscillogram was not substantially changed after denoising as much as it does with filtering (see Fig. 4.7) leading to a comparatively small MSE. However, these results altogether confirm that our wavelet denoising approach performs really well for birdsong. Even for the very low frequency kākāpō *booming* the denoising was still better with wavelets. On the other hand, in the case of less noisy bird calls, after denoising there was no significant information loss (see Fig. 4.7 (g)).

As discussed under Selecting the Best Decomposition Level, our automated method selects the appropriate decomposition level based on the complexity of the given signal. We classified the complexity of a song by examining the spectrogram pattern and listening to the sound; songs with more harmonics and wide frequency range were considered more complex (for example, in the case of ruru, trill calls are rather complex compared to narrow band more-pork calls). The results confirm that there is a relationship between the best decomposition level and the complexity of birdsong: if we order the calls according to their complexity from simplest to complex, the order is kākāpō *booming*, *more* and *pork*, kākāpō *chinging*, kiwi *female*, kiwi *male*, and finally ruru *trilling*, and this order can be seen in the depth of the tree (WMDL) in the last column of Table 4.3.

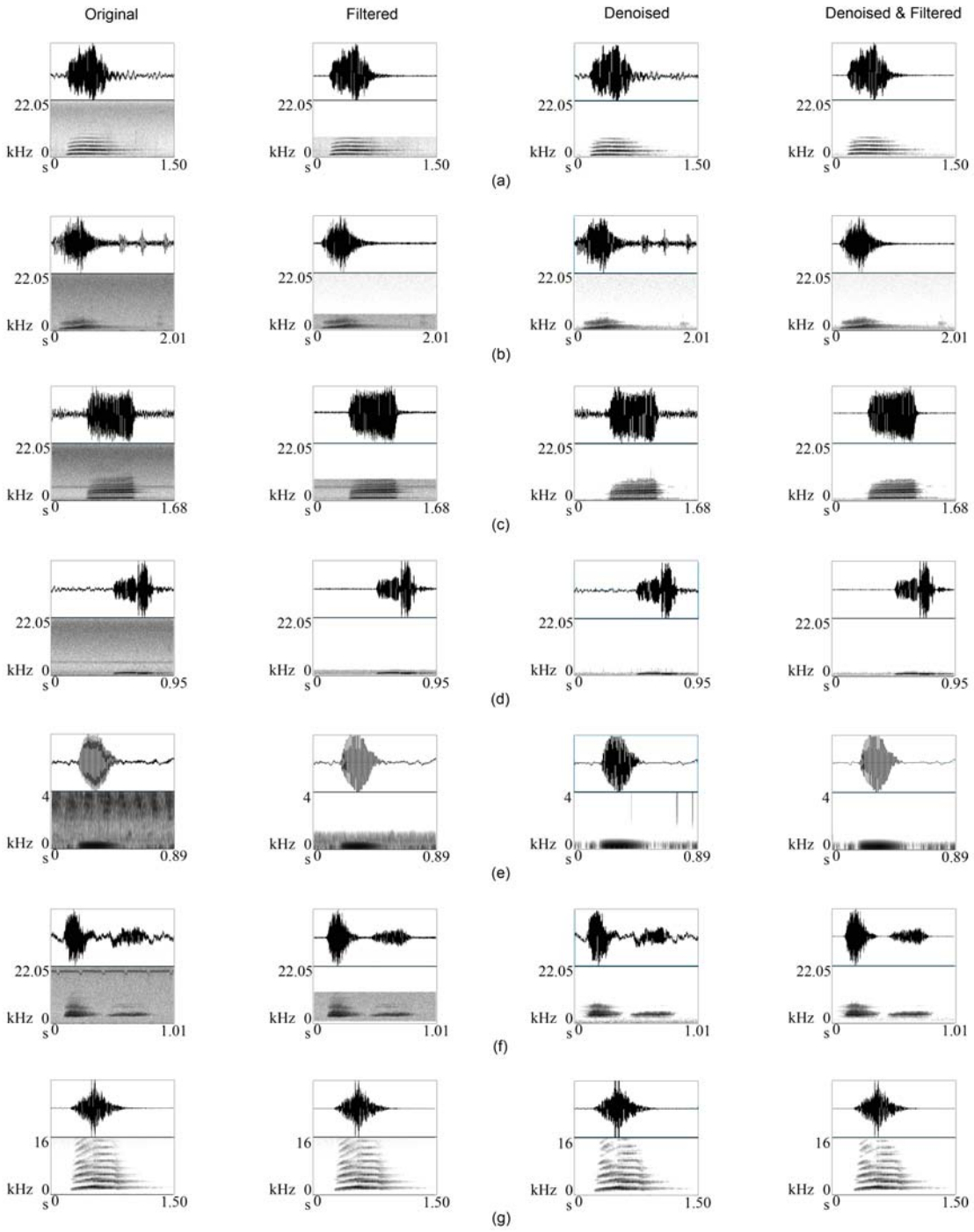


Figure 4.7: Bird call examples of before, after filtering, after denoising using wavelets as described in the text, and after denoising and classical filtering. (a) A whistle of a male North Island brown kiwi, (b) a call of female North Island brown kiwi, (c) a ruru *trill*, (d) a ruru *more*, (e) a kākāpō *booming*, (f) a kakapo *chinging*, and (g) a less noisy example of male kiwi.

Table 4.3: Experimental Results – primary dataset.

Species/		**O**	**F**			**D**			**DF**			**WMDL**
call type		SnNR	SnNR	S ratio	PSNR	SnNR	S ratio	PSNR	SnNR	S ratio	PSNR	
Kiwi	male	0.666	1.918	1.509	10.849	0.710	0.063	36.145	2.877	2.156	10.935	10
	female	0.405	1.379	1.536	11.860	0.423	0.035	39.633	1.690	1.801	11.942	10
Ruru	trill	0.341	0.988	1.075	10.902	0.365	0.261	22.569	4.792	3.840	11.695	10
	more	0.761	1.940	1.170	10.596	1.121	0.482	25.758	6.463	2.979	10.963	8
	pork	0.676	1.702	1.114	9.159	1.034	0.457	24.852	5.520	2.950	9.574	8
Kākāpō	boom	1.136	1.138	0.032	4.843	1.187	0.080	43.554	1.184	0.105	4.889	6
	ching	0.682	0.617	1.059	14.790	0.703	0.036	45.416	2.016	1.360	14.857	9
Total/mean		**0.667**	**1.526**	**1.071**	**10.428**	**0.792**	**0.202**	**33.990**	**3.506**	**2.170**	**10.694**	**9**

O=original calls. F=band-pass or low-pass filtered calls. D=wavelet denoised calls. DF=wavelet denoised and filtered calls. S ratio=success ratio, SnNR=Signal to Noise Ratio and PSNR=Peak Signal to Noise Ratio introduced in Evaluation Metrics. WMDL= Wavelet Mean Decomposition Level.

4.6.1 Extensions

Our method achieved impressive noise removal for the birdsong of the species we considered in the secondary dataset. Table 4.4 and Fig. 4.8 show that the overall SnNR reached more than seven fold (2.758) after the treatment compared to their initial SnNR (0.353). Some examples of these songs are presented in Fig. 4.9 (S 4.2 Audio).

While the main aim of our approach was to denoise individual bird calls, we also considered two extensions: denoising a series of bird calls in a sequence without segmenting them, and denoising a signal that is comprised of two or more overlapping bird calls. Table 4.5 and Fig. 4.10 compare the results of denoising unsegmented series of bird calls to their segmented calls. In the presence of unsegmented recordings, the mean SnNR of initial, filtered, and denoised songs were 0.548, 1.204, and 7.326 respectively. This success was confirmed by further analysis of their spectrograms and sound quality, for example, Fig. 4.9 (d) shows a denoised version of a series of kākāpō *chinging*. These results support the fact that denoising unsegmented long recordings is also possible and performs nearly equally to denoising their isolated calls.

The method worked very well even when presented with more than two overlapping birdsong, with the combination of the birdsong being retained, but the noise significantly reduced (Fig. 4.11, S 4.3 Audio). This is also reflected in the evaluation metrics where overall SnNR improved significantly from 0.556 to 5.222 (more than 9 times) after denoising and band-pass filtering compared to band-pass filtering alone (1.652 – less than 3 times) and denoising alone (0.634) for the ten examples described at the end of the Section Secondary Dataset. Confirming the potential, success ratio displayed a significant improvement after treating the examples with denoising and band-pass filtering (2.994) than filtering alone (1.261) and denoising alone (0.169). As usual, PSNR was highest with denoising alone (28.543) while filtering (12.818) and the combination of denoising and filtering (13.205) displayed relatively low PSNR because of the increase of MSE.

4.7 Discussion

The spectrogram has been the basis for much of birdsong analysis for decades, but it suffers from a fundamental tradeoff between temporal and frequency resolution because it is based on the Fourier transform. The more modern approach of wavelet analysis does not suffer from this tradeoff. However, there have been surprisingly few published studies in the field of birdsong recognition where wavelet analysis is used (Ren et al., 2008; Selin et al., 2007; Turunen et al., 2006)

Table 4.4: Experimental Results for the species introduced to the secondary dataset.

Species	**O**	**F**			**D**			**DF**		
	SnNR	SnNR	S ratio	PSNR	SnNR	S ratio	PSNR	SnNR	S ratio	PSNR
NI Robin	0.461	1.464	1.324	16.907	0.530	0.202	30.784	4.534	3.552	17.393
Tui	0.320	0.840	0.830	13.977	0.369	0.114	31.907	1.259	1.313	14.218
NI kākā	0.258	0.850	1.214	15.714	0.282	0.085	34.644	1.296	1.747	15.939
Hihi	0.651	1.814	0.745	11.446	1.133	0.692	29.259	3.217	1.884	11.794
NI saddleback	0.505	0.987	0.636	14.563	0.631	0.254	28.981	1.783	1.508	15.038
Marsh wren	0.220	1.910	2.426	16.511	0.219	0.011	41.837	3.423	3.280	16.579
Western meadowlark	0.180	2.277	2.520	12.437	0.181	0.011	41.807	2.906	2.994	12.474
Horned lark	0.225	2.271	2.491	12.356	0.228	0.017	39.846	3.647	3.448	12.402
Mean	**0.353**	**1.552**	**1.523**	**14.239**	**0.447**	**0.173**	**34.883**	**2.758**	**2.466**	**14.480**

Table 4.5: Comparing the denoising results – series of calls against their segmented calls.

Species/type		**#**	**O**	**F**			**D**			**DF**		
		e.g.	SnNR	SnNR	S ratio	PSNR	SnNR	S ratio	PSNR	SnNR	S ratio	PSNR
Kiwi (series)	male	5	0.584	1.015	0.522	13.325	0.968	0.513	23.453	12.159	3.610	14.172
	female	5	0.406	0.657	0.498	14.179	1.447	1.170	21.759	3.583	2.876	15.569
Kiwi (seg.)	male	20	0.659	1.187	0.585	11.588	1.359	0.626	20.776	14.128	4.365	12.456
	female	18	0.524	0.818	0.483	11.595	3.143	1.790	19.196	8.909	4.108	13.021
Ruru (series)	trill	5	0.272	1.269	1.334	13.647	0.288	0.207	25.504	16.453	4.828	14.337
	more-pork	5	0.290	1.143	1.689	13.125	0.566	0.375	26.587	7.782	4.088	13.386
Ruru (seg.)	trill	14	0.437	1.458	1.184	11.256	0.440	0.287	20.791	17.157	5.465	12.358
	more	11	0.464	1.413	1.449	10.337	1.367	0.716	22.830	5.547	3.433	10.689
	pork	11	0.308	1.209	1.457	9.157	0.584	0.808	20.865	4.459	4.122	10.040
Kākāpō (series)	boom	5	1.118	1.129	0.059	8.308	1.202	0.120	43.865	1.207	0.170	8.332
	ching	5	0.617	2.009	1.553	13.906	0.640	0.072	39.032	2.769	2.258	14.051
Kākāpō (seg.)	boom	21	1.184	1.196	0.063	4.810	1.245	0.082	43.037	1.250	0.135	4.838
	ching	17	0.768	2.069	1.467	12.286	0.802	0.076	38.627	2.709	2.062	12.426
Mean	**series**		**0.548**	**1.204**	**0.943**	**12.748**	**0.852**	**0.410**	**30.033**	**7.326**	**2.972**	**13.308**
Mean	**seg.**		**0.620**	**1.336**	**0.955**	**10.147**	**1.277**	**0.626**	**26.589**	**7.737**	**3.384**	**10.833**

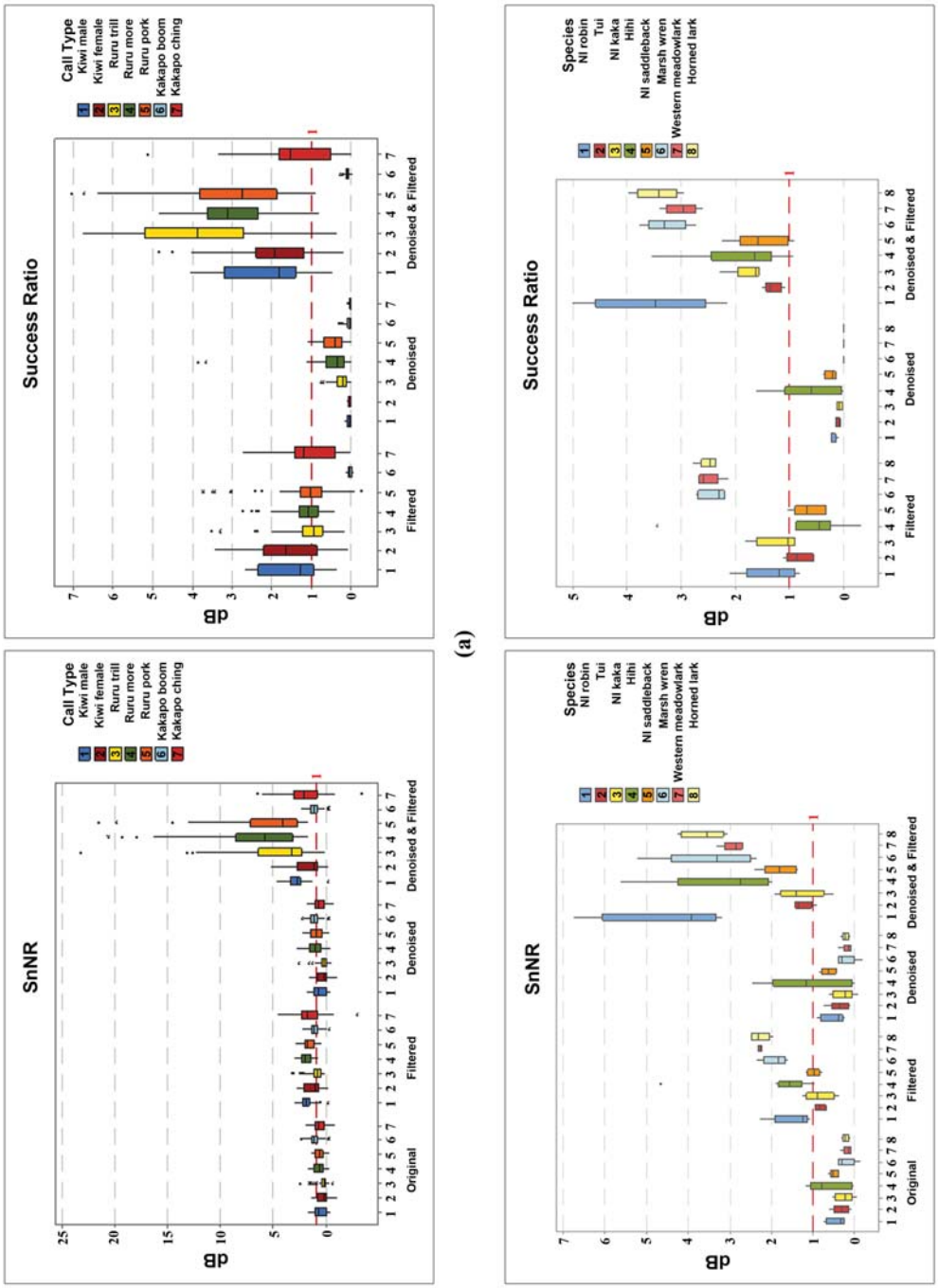


Figure 4.8: Box plot view of the results in (a) Table 4.3 and (b) Table 4.4.

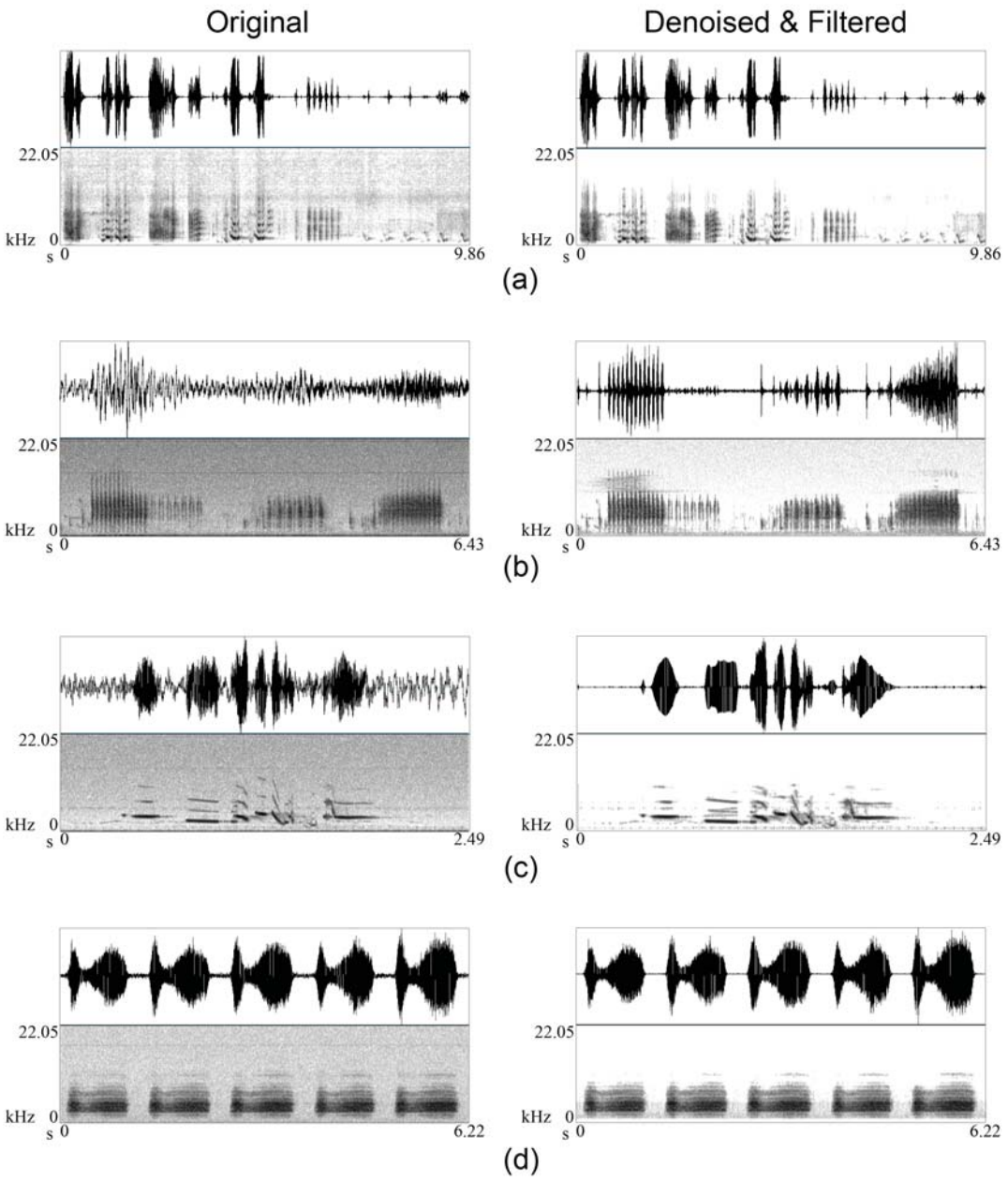


Figure 4.9: Denoising entire songs and long series of calls. (a) A North Island kākā song, (b) a marsh wren song, (c) a western meadowlark song, and (d) a series of kākāpō *chinging*.

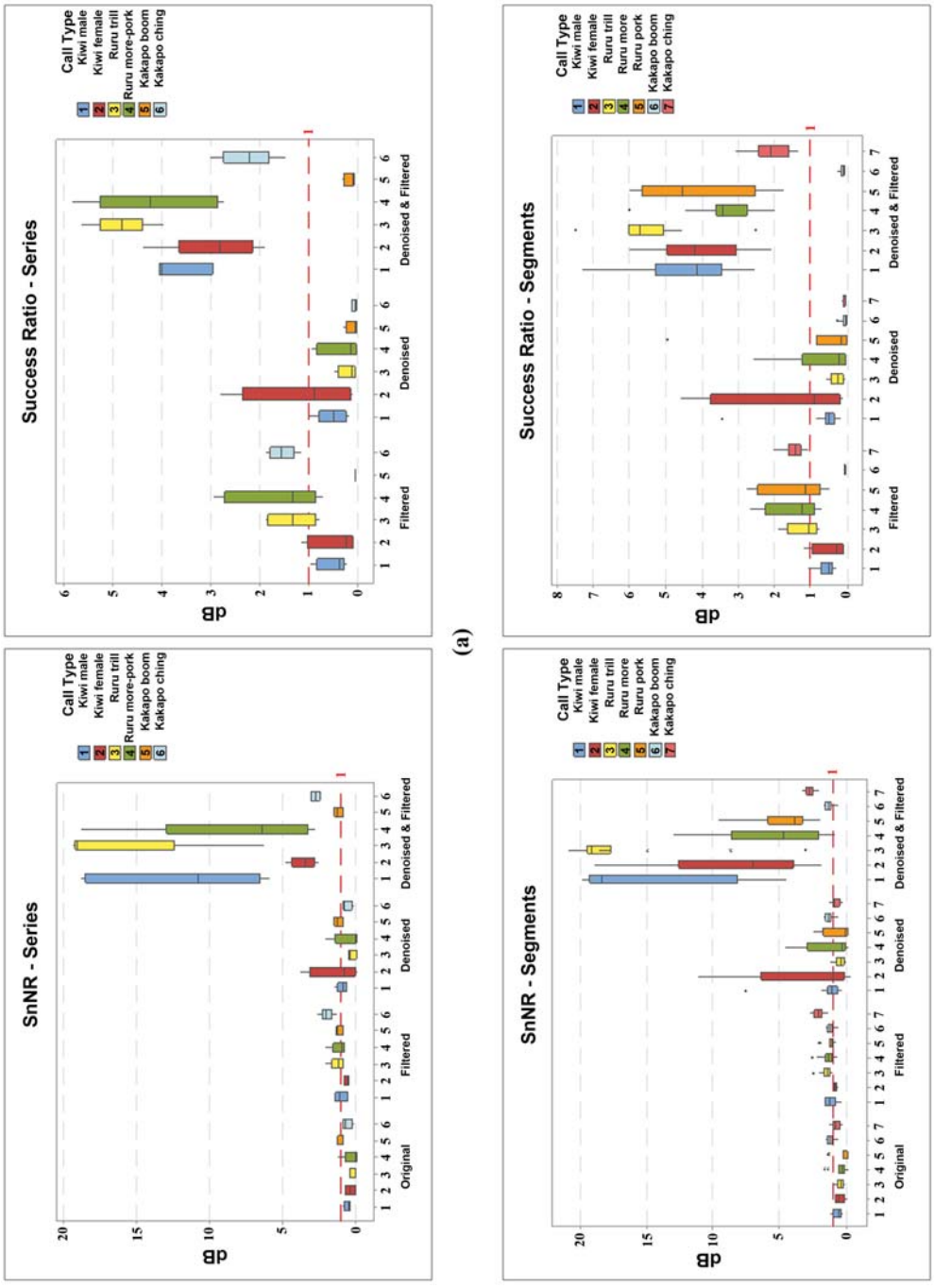


Figure 4.10: Box plot view of Table 4.5: (a) call series and (b) segmented calls.

In this paper we have investigated the use of wavelets for denoising automatically recorded birdsong. Denoising is becoming progressively more important as larger numbers of automatic recorders are deployed worldwide, recording not just birdsong, but every other noise in the environment. Whether these recordings are analysed automatically or manually, there is a need to reduce the extraneous noise from the recordings. We have demonstrated that wavelets are very good at washing out stationary noise from the recordings without distorting the birdsong. Even though some of the background noise (such as other animals) are not stationary, there is a substantial amount of stationary noise in recordings collected from nature. Therefore the applicability of this method to clean natural acoustic recordings is high. Further, much (although not all) natural noise is white or pink, and wavelets work well for removing it. Both the *success ratio* and modified SNR (SnNR) are useful measures of noise reduction. However, PSNR turned out to be a less reliable method to evaluate the success of noise reduction of audio signals.

This is one step towards our ultimate goal of automatically recognising birdsong by algorithm, and so our real aim is not the reproduction of a perfectly noiseless birdsong, but to remove the noise without damage to the signal, so that features of the song can be computed and used as input to other algorithms. In practice, the major reason for low recall rate or sensitivity (the percent of songs retrieved from the total number of songs in the recording) as well as low song recognition rate is the noise associated with the recordings: noise mixed with birdsong tends to hide song information (Swiston and Mennill, 2009; Buxton and Jones, 2012). Therefore, cleaning the recordings prior to call detection and segmentation would improve any method of song recognition. However, we have demonstrated that our method also allows impressive reproduction of denoised birdsong for use by biologists. On the other hand, this reproduction capability provides more flexibility to extract any preferred features for classification and recognition in contrast to the case in (Turunen et al., 2006).

The major challenge of using this method to clean long field recordings is its high computational cost: it requires significant computer memory and time. The complete analysis of approximately 2 minutes of calls (the segmented versions in Table 5) took approximately just over 10 minutes on a 2.4 GHz quad core i7. This can be improved through a compiled implementation of the method, rather than the general research-focused Matlab code that we have used here. In addition, we have demonstrated our approach on fairly long recordings so that we are confident that this method can be used to clean those too. Noise removal from the original recordings rather than from extracted isolated songs is really important both in semi-manual and in automated recognition. However, it is important to note that the level of noise, its nature, and strength of the song can cause significant effect when denoising using wavelets. For

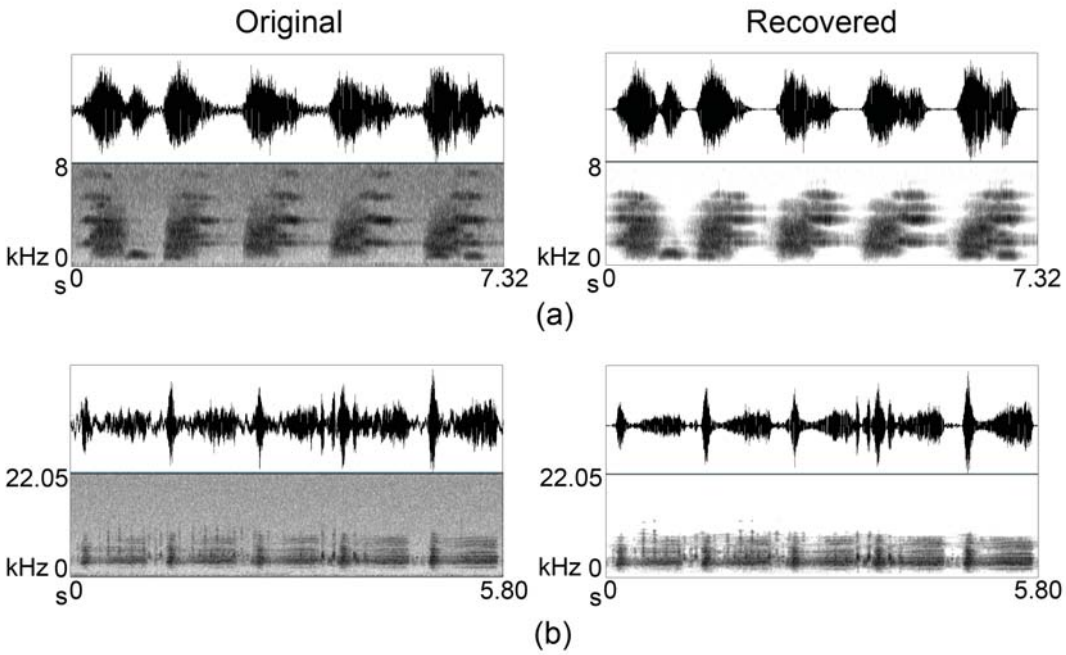


Figure 4.11: Denoising overlapped songs. Male kiwi, female kiwi, and more-pork are overlapped in (a) and kākāpō *chinging* overlapped with mottled petrels (*Pterodroma inexpectata*) in (b).

example we observed that denoising tended to remove both signal and noise when presented with very faded calls embedded in a high level of noise (calls that are hardly visible in the spectrogram). We observed the same when we inputted a North Island robin song mixed with strong noise at high frequencies (Fig 4.12, S 4.4 Audio). Interestingly, in this case, down-sampling saved the birdsong. If we initially apply low-pass filtering to filter out the frequencies beyond birds frequency range, we end up with a signal that contains the birdsong and non-Gaussian noise. This means that when we filter out the high frequency noise, the signal still has capacity for high frequencies unless we down sample it. Therefore, wavelet denoising cannot remove the remaining noise because it is non-Gaussian. In contrast, down sampling restricts the signal's frequency range, and automatically removes high frequency noise.

Generally, birds produce vocalisations within the range of human hearing. The dedicated recording devices we normally use in the field are also made to capture audible frequencies, but not ultrasonic ($>$ 20kHz) or infra-sound ($<$ 20Hz). However, many species produce sounds that are outside human and machine range. For example, species like the kākāpō, North African Houbara bustard (*Chlamydotis undulata undulata*), and bittern (*Botaurus lentiginosus*) generate *boomings* that are very low frequency signals (Cornec et al., 2014). These low frequency signals fall near or below

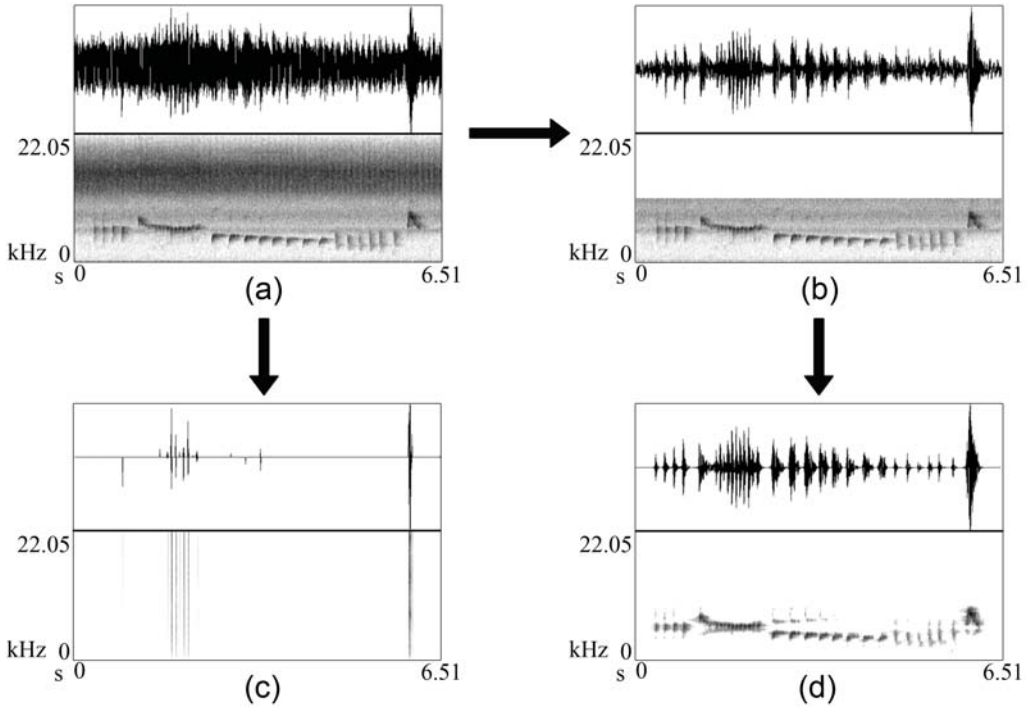


Figure 4.12: A deliberate denoising example. (a) A North Island robbin song (sampling frequency 44,100 Hz) and (b) its down-sampled song to 22,000 Hz. (c) and (d) are their denoised songs.

the threshold of human hearing (20 Hz). Birds have relatively greater hearing sensitivity than humans. For example, pigeons (*Columbidae*) have exceptional low-frequency (infrasound) perception (Hagstrum, 2000). However, current recording devices fail to fully capture these exceptional bird vocalisations. Accordingly, parallel to the development of birdsong recognisers there is a need of improving recorders and recording techniques.

In this study we have concentrated on birdsong, but automatic recorders also capture the sounds of other animals. In New Zealand for example, where introduced predators are responsible for the endangered status of many bird and reptile species, automatic recordings could be used to monitor populations of these introduced animals. Our denoising technique would work well to prepare recordings for identification and estimation of abundance of species such as stoats (*Mustela nivalis*), feral cats (*Felis catus*), rats (*Rattus spp*) and dogs (*Canis familiaris*) whose calls are high frequency.

Song detection from long recordings and segmentation is another sub-topic in the field of birdsong recognition, especially when it comes to practical use. The segmentation method used to isolate the bird songs has a huge influence on both the recognition rate and recall rate of a recogniser. Conventional energy based segmentations done using the waveform would easily skip faded songs in the recordings mainly as a result

of overlapping noise or the distance to the bird from the recording. On the other hand, this type of time domain approaches fail to separate bird songs from the background noise as they simply look at the energy and commonly a thresholding method to cutoff less energy sections. Therefore, this leads to increase false positives if the recogniser also fails to realise noise and discard them. But we speculate about using wavelet coefficients to do the segmentation in a more sophisticated manner. Separation of sound sources is another concept we did not consider in this context. Clearly it is not possible to separate sound sources easily in the presence of naturally recorded overlapping songs because the sounds are not linearly mixed even when we assume so, and the number of receivers (microphones) is always less than the number of sound sources. While the current study mainly focused on removing the stationary noise, it is essential to devise methods to tackle transient noise, but this would be more challenging because the birdsong are also transient.

Future work is to be carried out extending the usability of wavelets to address aforementioned gaps in this research area. We are currently investigating different feature extraction methods including MFCC (Priyadarshani et al., 2012) and wavelet coefficients as well as potential machine learning algorithms and similarity measures for recognition and classification of birdsong; it is important to determine the best combination of features that are strong enough to represent the birdsong uniquely. The final goal is to develop a non-species specific, robust and user friendly automated platform for ecologists to automatically process natural field recordings collected using any recorder.

4.8 Supporting Information

S 4.1 Audio. Bird call examples in Fig. 4.7

S 4.2 Audio. Birdsong examples in Fig. 4.9

S 4.3 Audio. Overlapped birdsong examples in Fig. 4.11

S 4.4 Audio. North Island robin song example in Fig 4.12

4.9 Acknowledgments

The authors would like to thank Bruce C. Robertson, Andrew Digby and the kākāpō recovery team for providing kākāpō recordings collected in 2009 and 2013, and Alex Brighten and Emma Mathison for providing ruru and North Island Brown kiwi calls. This paper was greatly improved by comments from Klaus Riede and Louis Ranjard.

Some of this work occurred while SM was at the Erwin Schrödinger Institute in Vienna, Austria.

Chapter 5

Birdsong Recognition in Continuous Field Recordings

Abstract

1. With one in every eight bird species around the world endangered, and many more being added to this list, monitoring of bird populations is a priority. Current acoustic monitoring methods are usually either too general: attempting to recognise a huge variety of species, many of which are not present in the area (or even country) of interest, resulting in lower accuracy, or too specific: being tailored to a single species and hard to retrain for other species. We need methods that are easy to train and use on a wide variety of species, without compromising on accuracy.

2. In this manuscript, we present a simple, yet scalable, method for detecting bird calls of a particular species from noisy field recordings. Rather than using filtering to select any sounds that stand out from the background followed by expensive recognition to discriminate the species or types of calls, we generate a filter that combines the segmentation and recognition phases and is specific to one species or one type of call. We evaluate the performance of the proposed detector with natural noisy field recordings of four species of birds recorded with automatic acoustic recording units and also an external dataset, including a species that is particularly hard to detect.

3. Our method had very high recall in detecting loud bird sounds (>95%) and even very faded bird sounds (around 70%) in continuous noisy field recordings. All the species gained >75% recall at <25% false positive rate. Using a playback and re-capture experiment we demonstrated that the proximity of the bird to the recorder has a strong effect on recording and consequently on detection.

4. High sensitivity and robustness of the proposed automated call detection method demonstrate that this method offers an unbiased and extremely efficient alternative to observer-based point counts and manual scanning of autonomous field recordings. The proposed method can be easily used for any species of interest, training can be done using even noisy field recordings, hence the applicability is high.

5.1 Introduction

In order to manage wildlife effectively it is necessary to have estimates of population abundance, since without this data it is impossible to determine the success or failure of management strategies. For bird populations, estimating the sizes of populations is difficult, as typically birds are hard to see, diffuse in the environment, and sometimes nocturnal. For these reasons, amongst others, bird vocalisation is the most effective way of surveying birds (Haselmayer and Quinn, 2000; Gregory et al., 2004) and call rates are often used to determine whether a bird population is stable, increasing, or decreasing over time (Colbourne and Digby, 2016; Dawson and Efford, 2009; Brandes, 2008).

This has given rise to the popularity of autonomous recording units, which make possible widespread and long-term recording of forest sounds in the absence of an observer. While the recordings are automatic, analysing them is still largely manual, and while there is a large amount of research into the general problem of birdsong recognition from their calls, methods that are accurate in the presence of noise and that can reliably detect birds that are far away from the microphone still elude us; unattended field recordings contain a great deal of environmental noise, and most of the currently available methods perform poorly with noise (Schrama et al., 2007; Wolf, 2009). Manual scanning of the recordings is time-consuming, requires a high level of expertise, is not scalable, and suffers from observer bias (Sauer et al., 1994; Emlen and DeJong, 1992). Of particular importance is the accurate recognition of the calls of endangered birds, since their populations are small, and thus misclassification is a significant issue. In addition, as the populations are small, it is also unlikely that many of the birds will be close to the microphone, and thus the accurate identification of the calls when they are significantly degraded by sound attenuation is crucial.

The first part of a recognition system is segmentation: taking hours of recording and extracting the syllables of birdsong (or other acoustic units) from it. In the presence of background noise this is far from trivial, and the majority of the published work has largely avoided segmentation and instead used manually segmented data (Somervuo et al., 2006; Anderson et al., 1996; Franzen and Gu, 2003; Fox et al., 2008) to test their recognition methods. Even when automated methods are used, they are often followed

by manual inspection (Selin et al., 2007).

The methods that have been developed for segmentation simply extract any areas of high acoustic energy from the recording. The extracted segments, or features derived from them, are then classified using standard techniques. These methods can be based on the amplitude of the signal or the frequency representation. Methods based on amplitude, such as Jinnai et al. (2012); Harma and Somervuo (2004); Somervuo et al. (2006); Towsey et al. (2012); Juang and Chen (2007), suffer from the fact that in unattended recordings the volume of the songs differ markedly as the noise is prominent and the bird sounds are often quiet (Briggs et al., 2012), since the proximity of the bird to the microphone varies.

Working in frequency space, one method that is commonly used and fairly successful is to treat the spectrogram as an image and applying image analysis techniques to detect regions of high power. The most common method is median clipping, where values greater than three times the median in each row and column of the spectrogram are marked as potential bird calls, and then morphological methods like dilation and erosion are used to form cohesive segments (Lasseck, 2013, 2015a,b; Potamitis, 2014; Fodor, 2013). This works well for calls that are clear, but does not detect calls that are quieter, nor work as well when the noise levels are high. It can, however, be used to detect the top and bottom frequencies of the call as well, which can be useful for further processing, although these will vary with distance as the harmonics of the signal are attenuated at different rates. Notwithstanding these problems, this form of segmentation has been used as a preprocessing stage successfully: Potamitis (2014) trained a random forest classifier with the features extracted from median clipping to recognise 78 bird species. He reached 91% area under the receiver operating characteristic curve (AUC; see Section 5.2 for details) on the test dataset. The dataset, however, included only short recordings (0.25–5.75 sec) and therefore the applicability of the method over long recordings has not yet been evaluated. Morphological opening (erosion and dilation) was employed by de Oliveira et al. (2015) to detect acoustic activity in 14 min long field recordings of Southern Lapwings (*Vanellus chilensis*) followed by a species-specific Hidden Markov Model (HMM) based recogniser that required large amount of training data (>27 h) to ultimately result in 56% recall despite the fact that very faint calls were excluded during the annotation.

Given the focus on detecting quiet bird calls in noisy recordings, we do not believe that energy-based methods will provide a complete answer to the problem in conservation. We therefore take a different approach to the problem. Rather than filtering for any sound that stands out from the background, we work through the recording (second-by-second), looking for particular combinations of frequency bands that are indicative of particular species of interest in each individual second of recording. These

can be detected reasonably reliably even in the presence of significant noise. The result is a filter that conflates the segmentation and recognition phases and is specific to one species or one type of call, rather than general: we are filtering out only (e.g.) kiwi calls from a recording, rather than any noticeable sounds, which could be produced by other species, other animals, aeroplanes, or something else again. Given a recording with many different types of bird call within it, using our method it will be necessary to process it several times, once for each species/type of call of interest. However, since natural autonomous recordings capture lots of other sounds that are not from birds, and since a separate recognition phase is not usually needed for our method (there may be some sets of calls that overlap and can be confused), we do not consider this a significant computational hardship.

In this paper we introduce our method, which is a simple learning algorithm based on wavelets, and compare it to two commonly used techniques: energy-based thresholding (Jinnai et al., 2012) and spectrogram-based median clipping (Lasseck, 2013, 2015b). We use recordings of four New Zealand birds of conservation interest: kākāpō (*Strigops habroptilus*), an extraordinary large flightless nocturnal parrot endemic to New Zealand which has only about 125 individuals left (Elliott, 2013); the cryptic Australasian bittern (*Botaurus poiciloptilus*) which has fewer than 1,000 individuals left in New Zealand (Williams, 2013); the brown kiwi (*Apteryx mantelli*) with about 25,000 individuals in four subspecies (Robertson, 2013) and the morepork (*Ninox novaeseelandiae*), New Zealand's only extant endemic owl. These birds produce a variety of sounds, some of which are extremely difficult to separate from noise in automatic song recognition, since they are so low pitched (the centre frequency of a bittern's call is around 150 Hz). However, there are no passerines amongst them. We therefore also used our method on a publicly available dataset of the American robin (*Turdus migratorius*), and compared our method to Potamitis et al. (2014) and Chu and Blumstein (2011), since they used that dataset.

We describe the datasets, performance measures, and our algorithm in the next section, before presenting the results and comparing our approach with two standard approaches to segmentation that are reported in the literature.

5.2 Materials and Methods

5.2.1 Datasets

Before assessing the suitability of the proposed method on real surveys, we used a playback experiment to explore the call detection efficacy at different distances between the bird and the recorder. Then we evaluated the method on unattended field recordings made in the natural habitats of four species of bird. The species are (at least partly)

nocturnal or crepuscular and some are highly camouflaged; therefore acoustic monitoring is key to estimating their populations. The species considered here produce a variety of sounds from very low frequency booming to high frequency whistles, representing different shapes in the spectrogram (Fig. 5.1). The very low frequency audio imprint of two of the species is hard to see in the spectrogram and is also not easily audible, because the sounds are hidden under low frequency noise in the environment such as wind and aeroplane noise. Finally, the proposed method was evaluated on an external dataset, the RMBL[1] Robin database. In all cases training data were kept completely separate from test data.

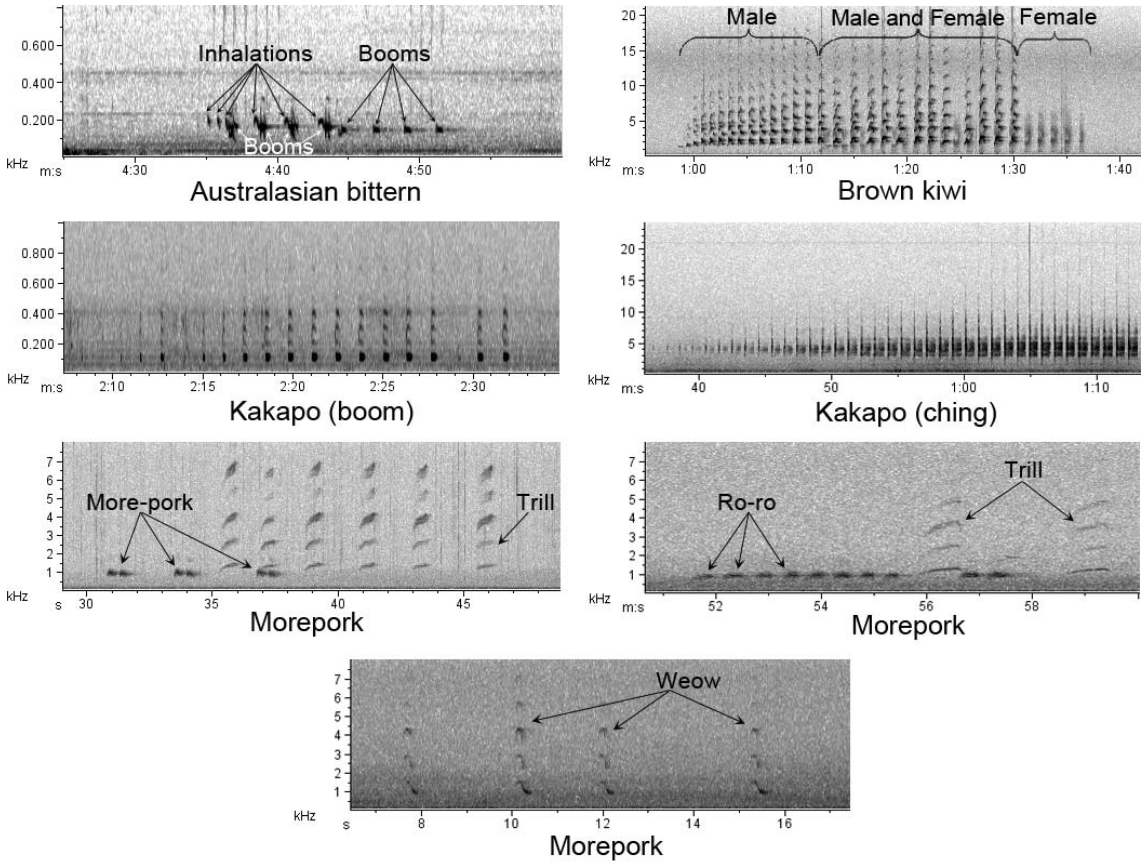


Figure 5.1: Close-range call excerpts from the unattended field recordings used to evaluate the proposed method. Spectrogram settings for Australasian bittern are given in (O'Donnell and Williams, 2015). Note that the scale of the vertical axis differs between the calls.

[1]http://www.seas.ucla.edu/spapl/projects/Bird.html

The Effect of Distance and Directionality of the Bird on Call Detection

A crucial aspect of detecting a call is how faded/quiet it is within the recording. The volume and direction of the call – as well as the detection limit of the recording device and microphones – will affect how a sound is recorded and later how a human or a computer can detect and recognise it. Despite a wealth of literature where recordings are used and analysed, we found no direct evaluation as to how sounds are lost with increasing distance from the recorder and how detectable them automatically. To fill this gap, we used a subset of a large dataset collected in other research (Chapter 3), to examine how call detection changes at increasing distance between the bird and the recorder.

The original experiment was carried out in a nearly flat dense forest, the Totara reserve (40°7′19.1″S 175°51′17.6″E) near Palmerston North, New Zealand. The site is located between a river and a road. Automated recorders were mounted on two perpendicular lines and a playback instrument (representing a bird) was placed at the intersection. The height of the recorders from the ground was approximately 1.5m (current practice by New Zealand wildlife managers is to mount the recorders at human ear level). Two transmission heights were set-up for the playback instrument: 0.25 m (for ground birds) and 3 m (for others). We generated the playback sounds at a volume as close as possible the actual sound of the particular bird species. Two speakers were used to transmit different bird sounds: a Boombox for very low frequency booming sounds, and a MiPro for all other sounds. For each sound sample, preliminary tests were carried out to tune the volume of the Sony PCM player to output the right volume through the speaker (Boombox or MiPro).

During the experiment, each bird sound (nine call examples representing the four bird species) was played back in four directions. Recorders were mounted at 20m, 50m, and 120m from the centre in each of four directions. We used three recorders mounted in one direction from the intersection (using bird direction as a variable) to test the effect of distance and the relative direction of the bird to the recorder. After the playback and re-capture session, from each recorder (three recorders), a five minute long file was created by merging eight re-recordings of nine bird sounds selected (bird facing the recorder, opposite, and 90°), including two repetitions. These three 5 min recordings were used to assess the capability of the recorder to capture the sound, and also to infer the capability of the detector to recognise calls at different distances between the bird and the recorder. Fig. 5.2 illustrates how a morepork call attenuated with increasing distance between the bird (speaker) and recorder and with the bird facing in different directions relative to the recorder. The attenuation with distance is clear, and at 120m most harmonics are lost and only the fundamental frequency of the bird call can be seen on the spectrogram. When the bird was facing the recorder relatively less attenuation

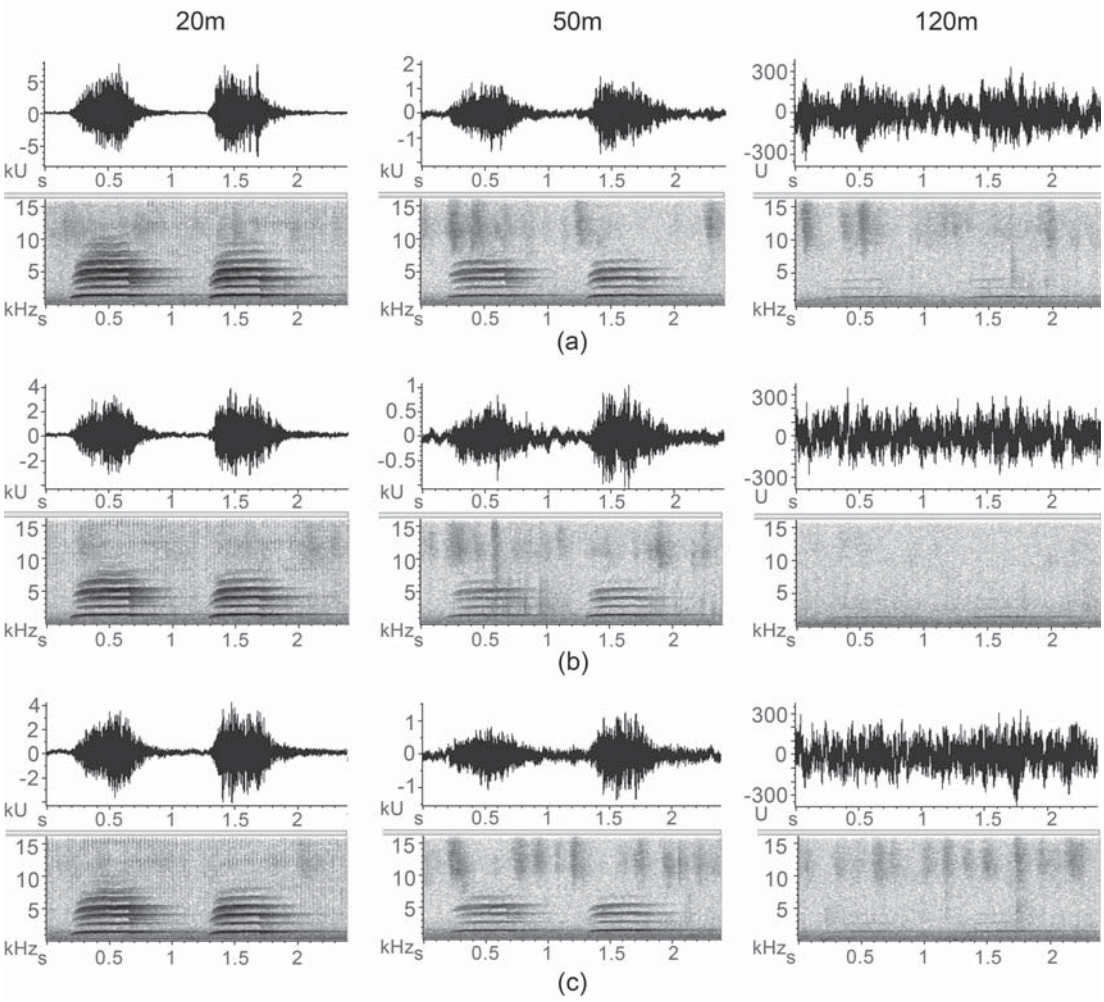


Figure 5.2: Re-recorded '*trill*' sound of morepork at different distances (columns) when the bird (speaker) was (a) facing the recorder, (b) at 90° to the recorder, and (c) facing away from the recorder.

was observed. It was also evident that the time domain signal (the waveform) is less useful for detecting bird sounds at moderate and long distance.

Species Used for Experiments

Species 1: Australasian Bittern

The Australasian bittern is an endangered swamp bird that is threatened throughout its range of New Zealand, southern Australia, and New Caledonia. While there is no information about rates of decline in New Zealand, current population estimations suggest that fewer than 1,000 birds remain (O'Donnell et al., 2013). In Australia, the population decline is above 38%, as reflected by the drastic decline in reporting rates over the last three decades (BirdLife International, 2016). The management of this species has

been very difficult because the bittern is an extremely cryptic bird with excellent camouflage, and so is very rarely seen. The New Zealand Department of Conservation has developed a protocol for the inventory and monitoring of Australasian bittern based on acoustic monitoring (O'Donnell and Williams, 2015). While females are almost silent, during the breeding season male bitterns produce very low frequency booms in a call sequence consisting of 1–10 individual booms (with a mean of 3) (Williams, 2013) and sometimes inhalations (Fig. 5.1). The centre frequency of bittern booms lies at 150 Hz and only the inhalations go up to 200 Hz (O'Donnell and Williams, 2015). Only the booms can be heard and seen in the spectrogram when the bird is not close to the recorder, therefore, passive monitoring of bitterns is mainly focused on the booming sound. In this study, however, we annotated both booms and inhalations recorded.

We used 15 minutes and 1 hour of recordings to train and test the bittern detector respectively. The recordings were collected using Department of Conservation acoustic recorders at Lake Whatuma, located in Waipukurau (40°1′26.79″S 176°31′20.61″E) New Zealand in 2010 and each call was labelled by an expert (Emma Williams) as part of a bittern monitoring program. We used that information plus careful spectrogram reading (O'Donnell and Williams, 2015) to generate precise annotation. Being very low frequency sounds, bittern booms in these uncontrolled recordings overlap with low frequency noise, e.g., aeroplanes, passing vehicles, and wind; there are also many sounds of other species in the recordings.

Species 2: Kākāpō

The kākāpō is an extraordinarly large nocturnal flightless parrot endemic to New Zealand, but unfortunately critically endangered. Predation by introduced mammals after human settlement brought them to the edge of extinction; the total population was 62 birds in 1991 (Waite et al., 2012) and with intensive management by the Department of Conservation the population was reported as 123 adults in 2016 (`http://kakaporecovery.org.nz/2016-media-releases`).

Male kākāpō produce deep low frequency booms and loud wheezing calls (named *chinging*) to attract mates (Fig. 5.1). The number of booms per booming bout varies among individual birds and ranges from approximately 8–16 (Harper, 1998). We observed that the fundamental frequency of booms lies approximately between 50–125 Hz, and it captures several harmonics up to about 500 Hz in close-range recordings. Chingings are bouts of particularly high frequency calls (approximately 2,000–10,000 Hz) produced in the vicinity of females. While booming and chinging are the main vocalisations of kākāpō we considered in this study, both males and females also produce high frequency *skraak* calls (Elliott, 2013).

As the two calls are very different, we trained two detectors, one for the booms and

one for the chinging. Individual training and testing datasets of 10 minutes and 30 minutes were used to assess the proposed kākāpō boom detector. The chinging dataset included 10 minutes of training data and 25 minutes of test data. The recordings include other bird species (e.g., petrels), insects, gusts of wind, aeroplanes, rain, thunder and other ambient noise.

Species 3: Brown Kiwi

Kiwi (*Apteryx spp.*), New Zealand's national bird, is a group of endemic nocturnal flightless birds. There are five species of kiwi, but the brown kiwi is the most abundant species in the wild. Kiwis can be more easily heard than seen due to their nocturnal behaviour. As shown in Fig. 5.1 brown kiwi males produce a sequence of high frequency whistles (15–25 whistles in a sequence) and females produce relatively low frequency cries (10–20 in a sequence) (Robertson, 2013). The frequency range of brown kiwi calls is approximately 500–8,000 Hz.

The frequencies of male and female brown kiwi overlap. We tailored the method to find both types of calls with a single detector. The detector was trained with 25 minutes of recordings and was tested with a different 80 minute recording. The dataset used here is a small subset of a large collection of autonomous recordings collected during 2013 by Alex Brighten (Brighten, 2015) and during 2014 by IC on Ponui Island (-36°51′59.99″S 175°10′60″E), New Zealand. Other species in the territory that co-vocalise with kiwi are mainly morepork, farm animals, and insects. Recordings also include high levels of abiotic noise (e.g., wind, rain, and aeroplanes).

Species 4: Morepork

Morepork, named after its distinctive *more-pork* call, is the only extant native owl in New Zealand. Presence of morepork is considered as an indicator of overall health of a forest. Moreporks are not considered threatened (Pohnke et al., 2015), but more attention needs to be paid, as they often consume human-introduced pest mammals e.g. mice that are the target of pest control efforts, meaning that they are at the risk of secondary poisoning (Brighten, 2015; Pohnke et al., 2015).

Morepork call throughout the year, and Brighten (2015) found that their repertoire consists of eleven call types and the frequency range of morepork is 500–11,000 Hz including harmonics (or up to 4,000Hz otherwise). Here we considered the common call types *more-pork*, *ro-ro-ro*, *trill*, and *weow* (Fig. 5.1), and made a single detector. Training and testing datasets comprised of distinct 10 minute and 45 minute recordings respectively. The datasets were from the same source as the kiwi dataset (Ponui Island) because kiwi and morepork co-exist in the same area.

All the above training and testing recordings were precisely annotated by NP in

discussion with IC and SM where necessary. Both training and testing recordings were treated in the same way during the annotation. More details about the methods for generating the ground truth annotations are given in the section describing the algorithm.

The RMBL Robin Database

In order to test our proposed method on a completely independent dataset annotated by a third party, we used an external dataset, the Rocky Mountain Biological Laboratory (RMBL) robin database of American robins made available by Chu and Blumstein (2011) (`http://www.seas.ucla.edu/spapl/projects/Bird.html`). The dataset includes 39 different length (1–7 minute) recordings collected from ten different locations (districts), annotated at the song and syllable levels in the software package Praat. The total length of the recordings is just over 78 minutes. Background noises include other species, human voice, wind, and the noise of water streams. Fig. 5.3 provides an example where the target species sings in the presence of another species in the background.

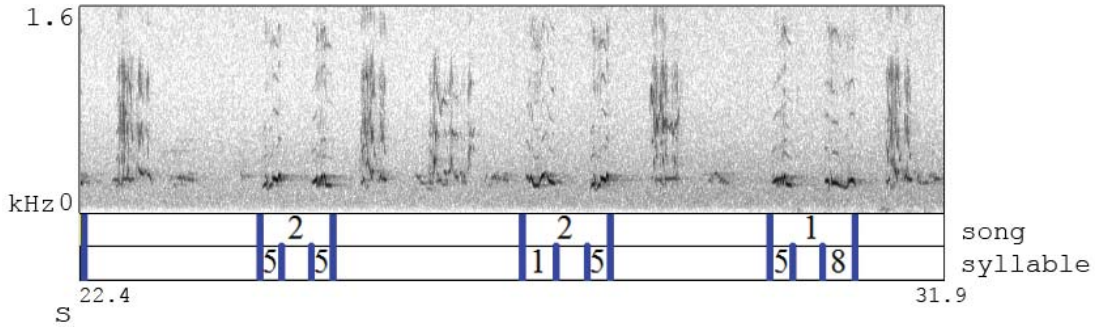


Figure 5.3: An excerpt from the RMBL dataset with its annotation. All American robin instances are labelled.

Success Measures

There are several measures that can be used to judge the effectiveness of a recogniser, although they are all based on the same four numbers: the true and false positives and negatives. From these numbers it is possible to compute *recall*: the fraction of relevant sounds that are identified (the true positive rate), *precision*: the fraction of the sounds positively identified as birdsong that were actually correct (true positives), and *specificity*: the true negative rate. When dealing with long autonomous field recordings, specificity is as important as recall and precision because the quantity of the recordings devoid of target species is usually larger than the amount with target vocalisations. Recall and precision can be seen as in opposition to each other: it is possible to gain

recall at the cost of losing precision and vice versa.

The final measure, which is the most commonly used, but is less useful, is *accuracy*, which is the proportion of the total number of predictions that were correct (both true positives and true negatives). One particular issue with accuracy as a measure is that when the two classes are unbalanced (there are typically far fewer positives than negatives) it does not reflect this. In acoustics, a low recall rate and low recognition accuracy are often caused by noise (Fox et al., 2006; Aide et al., 2013; Baker and Logue, 2003). While it is obviously important that the methods are as accurate as possible, the importance of the recall rate for rare species is worth noting, since for these birds confusing their sound with another (more common) species can lead to over-estimates of their abundance, or missing the fact that they exist at all in an area. Apart from the aforementioned four measures, their combinations are also commonly used to evaluate the performance. For example, the weighted average of recall and precision F_β (equation 5.1), the Receiver Operating Characteristic (ROC) curve, and the Area Under the ROC Curve (AUC). ROC visualises the performance of a binary classifier by plotting the True Positive Rate (TPR=recall) against the False Positive Rate (FPR=1-specificity) at various threshold settings, AUC summarises the plot. In particular, AUC is the probability that a classifier ranks a randomly chosen true positive higher than a randomly chosen false positive: an AUC of around 0.5 means that the classifier is not better than random guessing and an AUC of close to 1 means that the classifier is nearly perfect.

$$F_\beta = \frac{(1+\beta^2) \times (Recall \times Precision)}{(Recall + \beta^2 \times Precision)} \tag{5.1}$$

However, before any of these measures can be used, a decision needs to be taken as to how to measure the number of true and false positives and negatives. Comparing an automatically segmented recording with a manually produced ground truth can be done in several ways. One option would be to compare the start and stop times for each segment and give a score whenever these two times are within some pre-defined threshold of one another, and a negative score when they are not. However, instead we decided to score each second of the recording as to the presence or absence of a bird, and to count the number of seconds where the two annotations agree and disagree. This method is slightly biased towards detection, since any second that had both silence and birdcall in it would be marked as showing the presence of the bird. However, the unit of time can be decided appropriately e.g. half a second where the call length (syllable length) is relatively shorter.

5.2.2 Our Algorithm: Wavelet-Based Call Detection (Segmentation by Wavelet Filtering)

In a similar way to Fourier analysis, in the wavelet analysis a given signal is decomposed into a set of basis elements. However, rather than decomposing a signal into a set of sines and cosines of different frequencies (as in Fourier analysis), scaled and shifted versions of an irregular and compactly-supported basis function, namely the *mother wavelet* are used in the wavelet decomposition. These basis functions consequently allow temporal localisation of features, enabling wavelets to be used to analyse transient signals and signals with sharp changes. Therefore, wavelets provide a strong foundation for analysing bird sounds as we already saw (Chapter 4; Priyadarshani et al., 2016) when denoising bird recordings. In this paper we extend the performance of wavelets to detect target bird sounds from continuous unattended field recordings. A detailed mathematical foundation of wavelets is beyond the scope of this paper, but is readily available from a wide variety of papers, e.g. Daubechies (1992); Graps (1995).

Annotation

For training data we used a set of recordings that included the basic call variations from the species of interest. For most automatic identifiers to work, the training data need to be clean, but for our method, there is no such restriction: natural noisy field recordings can be used to train the automated detectors. However, target sounds in the training files need to be precisely labelled; the more accurate the annotation the more reliable the results, as the recogniser uses the labels to train the detector and later to identify those sounds in test recordings. The ground truth data for call presence within each time window in the training recordings was created by manually inspecting the spectrograms, combined with listening where necessary. We chose a one second resolution window to score the presence of calls in the recordings because we expected to have high resolution in the results, but the length of the time window is totally adjustable. Initially, we recorded the start and end of each call found (syllable level where appropriate), then converted it into an annotation that consisted of presence (1) or absence (0) for each of the L time intervals in the recording; at 1 second resolution, $L = 300$ for a 5 minute recording. The annotation vector for the training recordings is used by the training algorithm (Algorithm 1) to identify the wavelet nodes that represent the target bird sounds, as is explained below.

In order to test how the proposed method detects different *quality* birdsong, particularly the loudness or the completeness of the representation in the spectrogram, in real field recordings made in their natural habitats, we assigned a quality to each target bird vocalisation by visual, and sometimes acoustic, inspection during the annotation of the dataset. This quality label represents an indirect measure of the distance

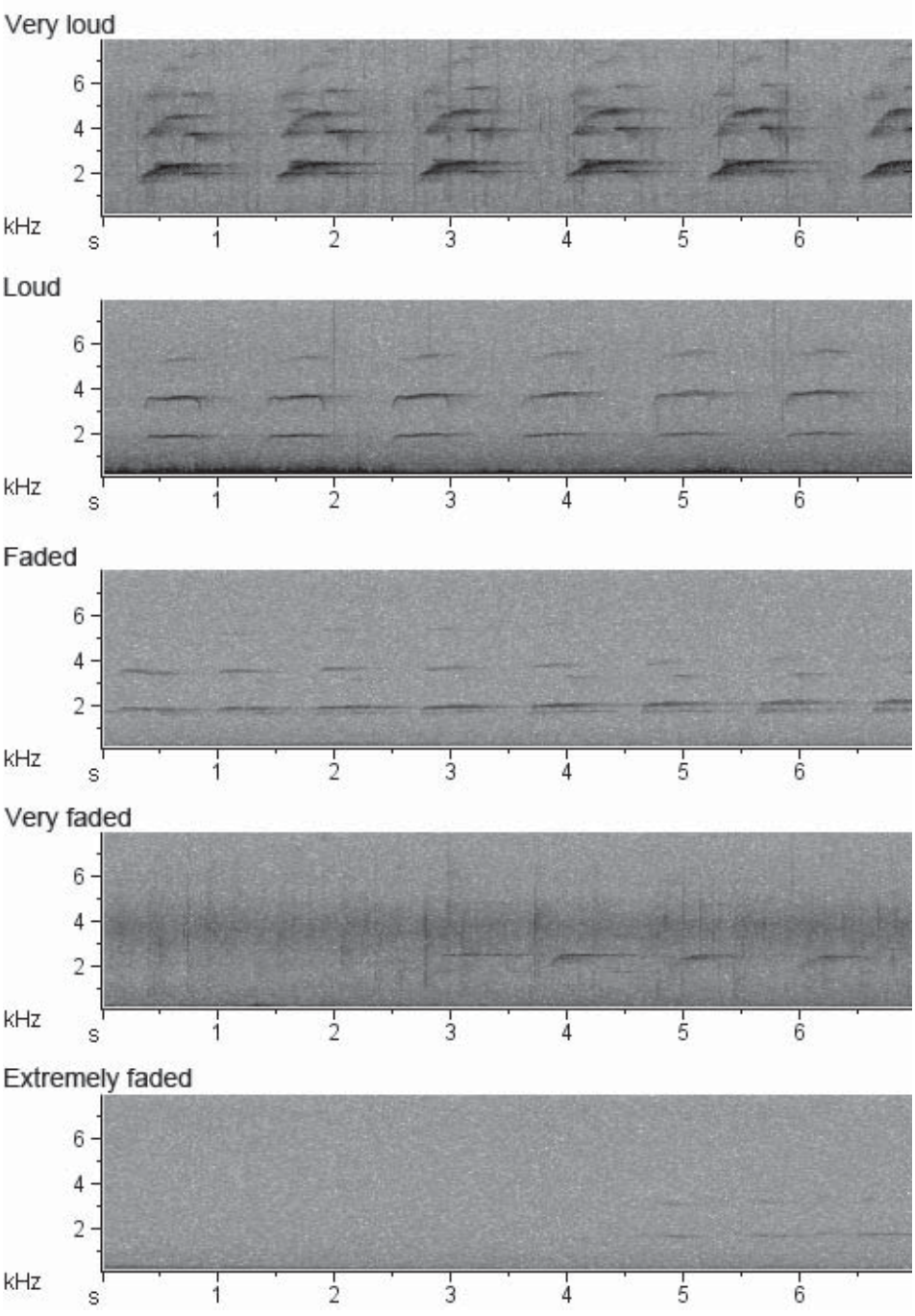


Figure 5.4: Call excerpts from the brown kiwi test dataset illustrating the different levels of quality of male whistles in field recordings. Note that all these examples were successfully detected by our kiwi detector.

of the bird to the recorder. The quality label ranged from 'very close' to 'extremely faded' (Fig. 5.4; S 5.1 Audio). We defined 'very close' to be a sound that contained the complete voiceprint of the recorded bird (including all harmonics where relevant); when the sound was loud but some harmonics were faded out, we classified it as 'close'; when the sound was faded and most of the harmonics were missing, we termed it 'faded'; when the sound was sufficiently degraded that only the fundamental frequency was visible we termed it 'very faded'; when the sound was barely visible or audible, it was considered 'extremely faded'. Effort was made to make this labelling as repeatable as possible, but it is a subjective method.

Training

The original recording was initially denoised as explained in (Chapter 4; Priyadarshani et al., 2016). Following that, the wavelet packet decomposition tree of the denoised recording was generated. Based on initial experiment the *discrete Meyer* wavelet (*dmey*) was chosen as the mother wavelet (Priyadarshani et al., 2016) and the decomposition was carried out to a depth of five, since this provided enough resolution to separate the noise and signal. In the example shown in Fig. 5.5 the frequency width of leaf nodes is small (approximately 16 Hz for bittern). In total there were 62 nodes in the tree excluding the root. The original data are spread over these nodes, although relatively few nodes contain most of the energy of the signal. Accordingly, the aim was to identify the wavelet nodes that are sensitive to target bird species, but not to the other sounds present in the recording.

The energy coefficients for each node corresponding to each second of the recording were placed in a matrix $E_{N\times L}$, where N is the number of nodes (62) and L is the number of time intervals in the recording, as previously. Each row of the matrix represents the energy of the particular node for every second in the recording. The ground truth annotations stating presence or absence were stored in another matrix $A_{L\times 1}$. The point-biserial correlation between each node and the annotation matrix ($C_{N\times 1} = E_{N\times L} \cdot A_{L\times 1}$) was used to find which nodes better represent the target species. The correlation matrix $C_{N\times 1}$ represents the correlation of each node to the ground truth. High positive correlation means that a particular node represents the signal more than the other nodes, while negative correlations imply that those nodes have captured noise.

The number of nodes to be considered depends on the complexity and the frequency range of the species being considered. Preliminary trials confirmed that considering nodes beyond the top twenty correlated nodes was useless. The nodes that represent sounds in the frequency band of the target species are listed in order of their correlation to the ground truth and the level of the node in the tree. Initially, the node with the

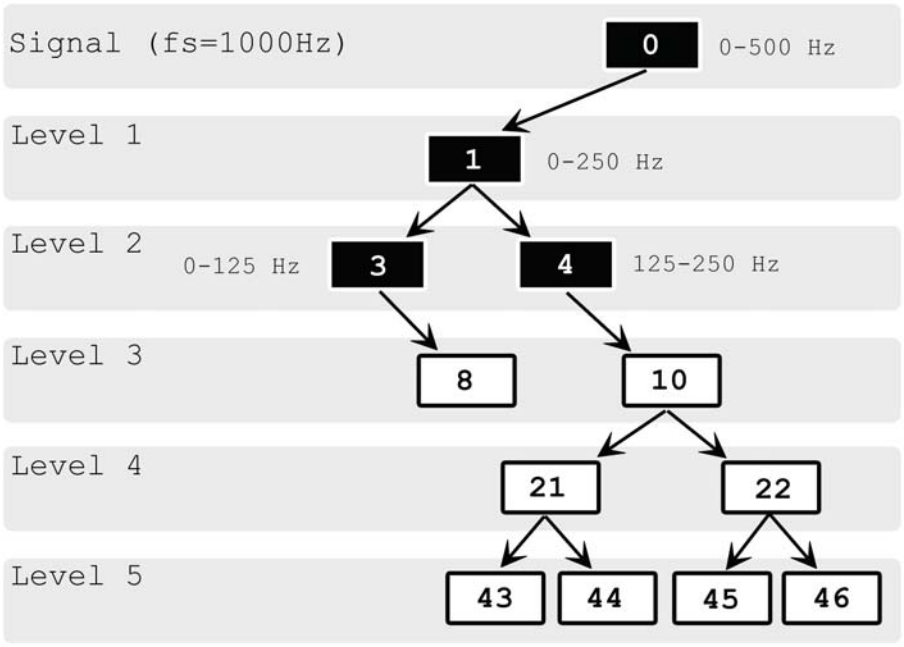


Figure 5.5: Optimised wavelet packet decomposition tree for bittern detection. The white nodes are the ones that are sensitive to bittern booms. The filtered nodes occupy 62–250 Hz.

Table 5.1: The three training examples (*Train1*, *Train2*, and *Train3*; length of each example is 5 min) used to train the bittern detector. The nodes that contributed to improved F_2 score were filtered, e.g. nodes 44, 46, 22, 21, and 10 in the case of *Train1*.

Train1	Highest correlated nodes in descending order of correlation [4, 10, 44, 46, 22, 21, 41, 20, 9, 45]									
Reordered nodes	**44**	**46**	41	45	**22**	**21**	20	**10**	9	4
Recall	0.53	0.60	0.60	0.60	0.70	0.75	0.75	0.86	0.86	0.86
Precision	0.79	0.79	0.77	0.77	0.80	0.81	0.81	0.80	0.80	0.80
F_2	0.56	0.63	0.63	0.63	0.72	0.77	0.77	0.85	0.85	0.85
Train2	Highest correlated nodes in descending order of correlation [43, 35, 21, 3, 17, 9, 8, 41, 20, 2]									
Reordered nodes	**43**	35	**21**	17	**8**	3	41	20	9	2
Recall	0.66	0.69	0.84	0.84	0.93	0.93	0.93	0.93	0.93	0.93
Precision	0.41	0.37	0.33	0.32	0.35	0.35	0.35	0.35	0.35	0.35
F_2	0.59	0.59	0.64	0.64	0.70	0.70	0.70	0.70	0.70	0.70
Train3	Highest correlated nodes in descending order of correlation [41, 20, 45, 9, 22, 2, 4, 6, 46, 5]									
Reordered nodes	41	20	**45**	9	**46**	**22**	6	5	2	4
Recall	0.00	0.00	0.23	0.23	0.46	0.69	0.69	0.69	0.69	0.69
Precision	–	–	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
F_2	–	–	0.27	0.27	0.52	0.74	0.74	0.74	0.74	0.74

Algorithm 1 Segmentation by Wavelet Filtering: Training

Require: S, $A_{L\times 1}$ ▷ S - sound file (sampling frequency - f_s), $A_{L\times 1}$ - annotation
Ensure: LN ▷ LN - list of nodes
  if $f_s \neq f_{pre}$ **then** ▷ f_{pre} - preferred sampling frequency
    Subsample S
  $S \leftarrow$ denoised S (dmey, ≤ 15 levels)
  $S \leftarrow$ band-pass S (given frequency range)
  $T \leftarrow$ wavelet packet decomposition of S (dmey, 5 levels)
  $E_{N\times L} \leftarrow$ energy in each node in the tree T
  $C_{N\times 1} \leftarrow$ point-biserial correlation between $E_{N\times L}$ and $A_{L\times 1}$
  $Nodes \leftarrow$ highest correlated 20 nodes (out of 62)
  $Nodes \leftarrow$ re-sorted $Nodes$ by tree order and the correlation (so that children nodes are before their parent nodes)
  $LN \leftarrow \{\}$
  for $node \in Nodes$ **do**
    if F_2 increases with node $node$ **then**
      $LN \leftarrow \{LN, node\}$

highest correlation is selected, if it is a leaf node, then it comes first in the node list. Otherwise any children of that node in the list (if any) are given priority (with respect to the correlation value when more than one of them is present) and then the method recurses to the next level of the tree. This process is applied to all the nodes in the initial list to generate the new sorted list. For example, consider the first training instance of bittern (*Train1*) in Table 5.1. There, the highest correlated node (4) is a non-leaf node (Fig. 5.5), therefore after rearranging them according to our rule, its leaf nodes 44, 46, 41, 45, and then their parent nodes (one level up) 22, 21, 20, and then their parent nodes 10 and 9 got priority.

From this ordered list we then choose nodes to include in the classifier by starting from the first node in the list and including nodes that improved the classification performance, as measured by the F_2 score (equation 5.1), which combines recall and precision. A node is included to the optimum node list only if it contributed to improving the F_2 score (see Table 5.1 and Fig. 5.5).

During call detection, a smooth energy curve was generated over the wavelet coefficients of each node, and then a threshold was applied to detect target sounds. The energy curve was generated as explained in (Jinnai et al., 2012), but replacing the amplitude by wavelet coefficients (without normalising the energy curve periodically). Rather than deciding the threshold for each example manually, we used *mean + standard deviation* threshold calculated over the wavelet coefficients in all cases except in the case of bittern. For bittern calls the *mean + standard deviation* did not provide good recall and we used a ratio of *minimax*. The threshold calculated this way was comparable to those calculated with *mean + standard deviation* for the other species

calls. The sections of the recording where the energy exceeded the threshold were considered as target calls, as shown in Fig. 5.6 (S 5.2 Audio).

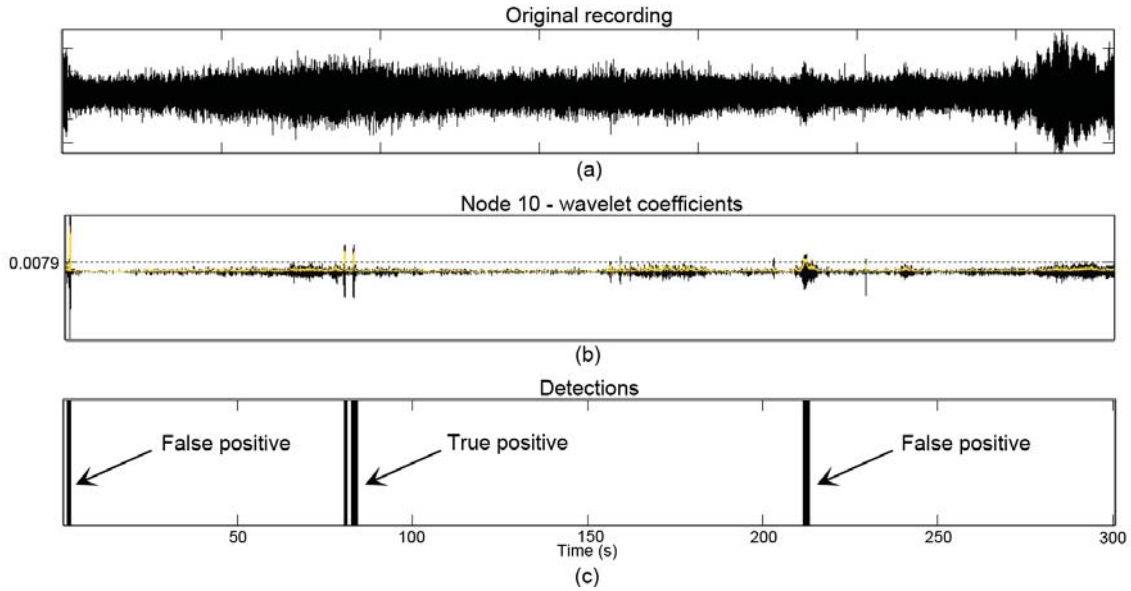


Figure 5.6: Detection of bittern calls using a wavelet node from (a) the first 5 min of the test dataset (Table 5.4). (b) The energy curve (shown in yellow) generated over the wavelet coefficients (node 10) from (a) and (c) the binary output of call availability.

In the case of low frequency booms, the recordings (original sampling rate of 44,100 Hz) were re-sampled to 1,000 Hz prior to training and testing. During the training, only the top ten correlated nodes were useful. The nodes that were sensitive to bittern and kākāpō booms were [8, 10, 21, 22, 43, 44, 45, 46] and [3, 8, 17, 18, 36, 38] respectively.

For all other cases, the sampling frequency was set to 16,000 Hz and re-sampling was applied where necessary. Even though we focused on the first 20 nodes (e.g. due to the wider frequency range), it turned out that the optimum nodes were always found within the first 10 nodes. The wavelet filtering algorithm produced the node sets [33, 37, 38] and [20, 42, 55] for morepork and kākāpō chinging respectively. The algorithm found that the nodes [34, 35, 36, 38, 40, 41, 42, 43, 44, 45, 46, 55] represented kiwi. The smoothed energy curve was calculated on the reconstructed signal from each node filtered.

After finding the optimum node set within the training data, the detection algorithm (Algorithm 2) was guided by those nodes, was stable for the particular species, and was ready to test/use. Detections generated from each node (e.g. Fig. 5.6) were aggregated to form the final output.

Algorithm 2 Segmentation by Wavelet Filtering: Testing

Require: S, LN ▷ S - sound file (sampling frequency - f_s), LN - optimum node list
Ensure: Out ▷ binary output - presence/absence of target species in each time unit
if $f_s \neq f_{pre}$ **then** ▷ f_{pre} - preferred sampling frequency
Subsample S
$S \leftarrow$ denoised S (dmey, ≤ 15 levels)
$S \leftarrow$ band-pass S (given frequency range)
$T \leftarrow$ wavelet packet decomposition of S (dmey, 5 levels)
for $node \in LN$ **do**
Generate binary output Out_{node} with respect to $node$ ▷ call presence/absence in each time unit (1 sec)
Combine the outputs (Out_{node}) using OR operator to generate Out

5.2.3 Comparison Between Segmentation by Wavelet Filtering and Other Methods

Our method was compared to two commonly used methods: time domain energy thresholding (Jinnai et al., 2012) and spectrogram-based median clipping (Lasseck, 2013, 2015b). We used the same test dataset used for evaluating our wavelet call detection algorithm in those other methods using the author's guidelines and calculated recall, precision, specificity, and accuracy. Band-pass filters and resampling were employed (to avoid frequencies outside the bird's frequency range) in the same way as we used them in the proposed wavelet filtering algorithm. A threshold of 0.7 was applied on the normalised energy curve (to capture the loudest 70%). In the spectrogram-based method, values greater than three times the median in each row and column of the spectrogram were set to 1 and others to 0 producing a binary image of the spectrogram. The minimum blob area of 300 pixels turned out to be useful to extract segments after median clipping followed by closing (diamond shape with size 3), dilation (square shape size 2), and median filter. Note that both these methods will detect any sound, not just the relevant bird sounds. Therefore, after detecting the sounds, usually a classifier is used to filter the target bird sounds. This means that the comparison to our method is not fair, as our algorithm is performing both detection and classification simultaneously.

5.3 Results

5.3.1 The Effect of Distance and Direction of Bird on Song Capture and Automatic Detection

When the bird (playback) was close to the recorder (20m), the overall recall rate was very high (90%). But, as we expected, the recall rate decreased with distance to 88% at 50m, and to 71% at 120m. It was confirmed that when the bird was facing the

recorders, the recorders captured most of the signal and hence the recall was highest (91%). Unexpectedly, the recall was the same (88%) when the bird was facing the opposite direction to the microphone and when at 90° to the recorder.

Even though evaluating the proposed wavelet filter is not the primary objective of this playback and re-capture dataset, we present the results in Table 5.2. Note that the recorders also captured morepork vocalisations made by the birds present at the experiment site during our playback-recapture session. Those non-playback instances were eliminated from the above recall rate calculations because we did not know how far away the birds were when producing the calls and in which direction were they facing. These calls were however included in precision, specificity, and accuracy measures.

In detail, the kiwi wavelet filter could detect all the brown kiwi calls recorded by the first two recorders (20m and 50m) regardless of the direction of the bird, and also all the brown kiwi calls recorded when the bird was pointing to the recorder at 120m. When the bird was 120m away from the recorder facing the opposite direction and 90° to the recorder, the recall of the detector reduced to 43% and 59% respectively. The precision of the kiwi wavelet call detection algorithm was only slightly reduced with distance (62% at 20m to 58% at 120m) and the accuracy was always above 77%.

The morepork wavelet filter successfully detected all the morepork calls recorded while the bird was facing the recorders regardless of the distance (20m–120m range). The recall was slightly reduced when the bird was calling in the opposite direction to the recorder (20m – 94%, 50m – 94%, and 120m – 88%) followed by 90° to the recorder (20m – 82%, 50m – 79%, and 120m – 79%). Consistently, the morepork detector maintained the precision, specificity, and accuracy above 67%, 82%, and 80% respectively.

The broadcasted bittern booms/inhalations and kākāpō booms were not audible to the recorder (or visible in the spectrogram) at 120m and when the bird was pointing in the opposite direction to the recorder at 50m. At 20m, the highest bittern recall was found when the bird was facing the recorder (77%), followed by facing 90° degrees to the recorder (75%) and facing opposite to the recorder (58%). The precision at 20m was 78%. The bittern wavelet filter failed to detect a few bittern booms captured by the recorder at 50m. Specificity and overall accuracy were high and improved with distance: 20m – 96%, 50m – 100%, 120m – 100% precision and 20m – 92%, 50m – 97%, 120m – 100% accuracy.

All the kākāpō booms were successfully detected by the kākāpō wavelet filter (boom) at 20m, the precision was low (15%) but the specificity and the accuracy were around 70%. Deviating from the aforementioned overall results, the recall was higher when the bird was at 90° (83%) compared to when the bird was facing the recorder (67%) at 50m. Both specificity and accuracy settled at 69% at 120m.

Table 5.2: The effect of the distance and the direction of the bird to the recorder in call detection of each species considered. Values are the number of seconds where calls were detected per recording by the wavelet filtering algorithm. These values are contrasted to careful annotation of the same recordings by a human expert to obtain the following values: TP=true positives, FP=false positives, TN=true negatives, and FN=false negatives. All calls were playbacks with the exception of some morepork calls produced by wild birds at the experimental site. R=direction of the recorder and O=opposite direction.

Distance to recorder from speaker			**20m**			**50m**			**120m**		
Speaker facing			**R**	**90°**	**O**	**R**	**90°**	**O**	**R**	**90°**	**O**
	Species	**Call type**									
TP (in sec)	**bittern**	*boom*	9	17	7	0	0	0	0	0	0
		inhalation	1	1	0	0	0	0	0	0	0
	kākāpō	*boom*	3	8	4	2	5	0	0	0	0
		ching	8	15	7	8	14	6	1	0	0
	brown kiwi	*male*	15	31	15	15	31	15	15	14	3
		female	6	10	6	6	10	6	6	10	6
	morepork	*more-pork*	7	12	6	7	12	6	7	11	6
		trill	9	16	9	9	15	9	9	15	8
		wild morepork in background	7			17			30		
FN (in sec)	**bittern**	*boom*	0	1	1	3	6	0	0	0	0
		inhalation	3	5	4	0	0	0	0	0	0
	kākāpō	*boom*	0	0	0	1	1	0	0	0	0
		ching	0	1	2	0	2	3	7	0	0
	brown kiwi	*male*	0	0	0	0	0	0	0	17	12
		female	0	0	0	0	0	0	0	0	0
	morepork	*more-pork*	0	0	0	0	0	0	0	0	0
		trill	0	6	1	0	7	1	0	7	2
		wild morepork in background	22			13			9		
Recall			95%	89%	87%	92%	84%	91%	84%	68%	62%
FP (in sec)	**bittern**		10			0			0		
	kākāpō	*boom*	87			105			94		
		ching	97			106			95		
	brown kiwi		51			50			39		
	morepork		32			36			35		
TN (in sec)	**bittern**		241			291			300		
	kākāpō	*boom*	198			186			206		
		ching	170			161			197		
	brown kiwi		166			167			178		
	morepork		173			168			161		
Precision			45%			39%			35%		
Specificity			77%			77%			80%		
Accuracy			78%			78%			79%		

When the bird was calling towards the recorder, the kākāpō chinging detector was successful to detect all the chinging recorded by two recorders at 20m and 50m distance, but only 13% were recorded by the recorder at 120m. Recall was better when the bird was at 90° to the recorder (20m – 94% and 50m – 88%) than facing opposite to the recorders (20m – 78% and 50m – 67%). Precision was low (24%) but the specificity and accuracy were above 60%. The furthest recorder did not capture any kākāpō chinging when the bird was calling the opposite direction or 90° to the recorder.

In summary, when examining all the species, while specificity and accuracy were maintained around 78% the precision was low at 40%. This was mainly caused by the low precision obtained for kākāpō calls. All the kiwi and morepork playbacks were heard (captured) by the recorder at 120 m, but the kākāpō and bittern booms (and most of the kakapo chinging) were not captured (Table 5.2).

5.3.2 Results on Non-Experimental Continuous Field Recordings

ROC graphs for each wavelet filter are presented in Fig. 5.7. The horizontal and the vertical axes represent False Positive Rate (FPR=1-specificity) and True Positive Rate (TPR=recall) respectively. These graphs were generated by leaving the threshold variable.

Brown kiwi had the highest AUC (0.94) followed by kākāpō ching (0.89), morepork (0.86), Australasian bittern (0.83), and kākāpō boom (0.82). A similar pattern in ROC curves can be seen except in the case of Australasian bittern, in which the recall (TPR) was considerably low at higher specificity (lower FPR) but rapidly improved approximately after about 0.10 FPR. In contrast, the wavelet filter of kākāpō booms demonstrated higher recall at higher specificity and did not improve much after about 0.10 FPR.

Table 5.3 summarises the performance of each detector on natural noisy field recordings using the threshold given in the previous section. This table reflects a single point of each ROC curve in Fig. 5.7 which matches the selected threshold. The wavelet filtering algorithm achieved more than 95% recall in detecting close-range calls ('very loud' and 'loud'). Even when the calls were very faded the recall was just below 70%. The detector was successful in detecting 30% of extremely faded calls. In the case of bittern and kākāpō booming calls we had to revisit the human ground truth labels after using our algorithm: a few extra calls were found by our algorithm, and after retrospective consideration, included to the annotation. These newly detected booms were so faded that it was very difficult to observe them in the spectrogram.

Table 5.4 provides the overall and species level recall, precision, specificity, and accuracy measures on the same dataset in Table 5.3 with the same threshold settings. The overall accuracy of the detector, the accuracy of accepting target bird sounds

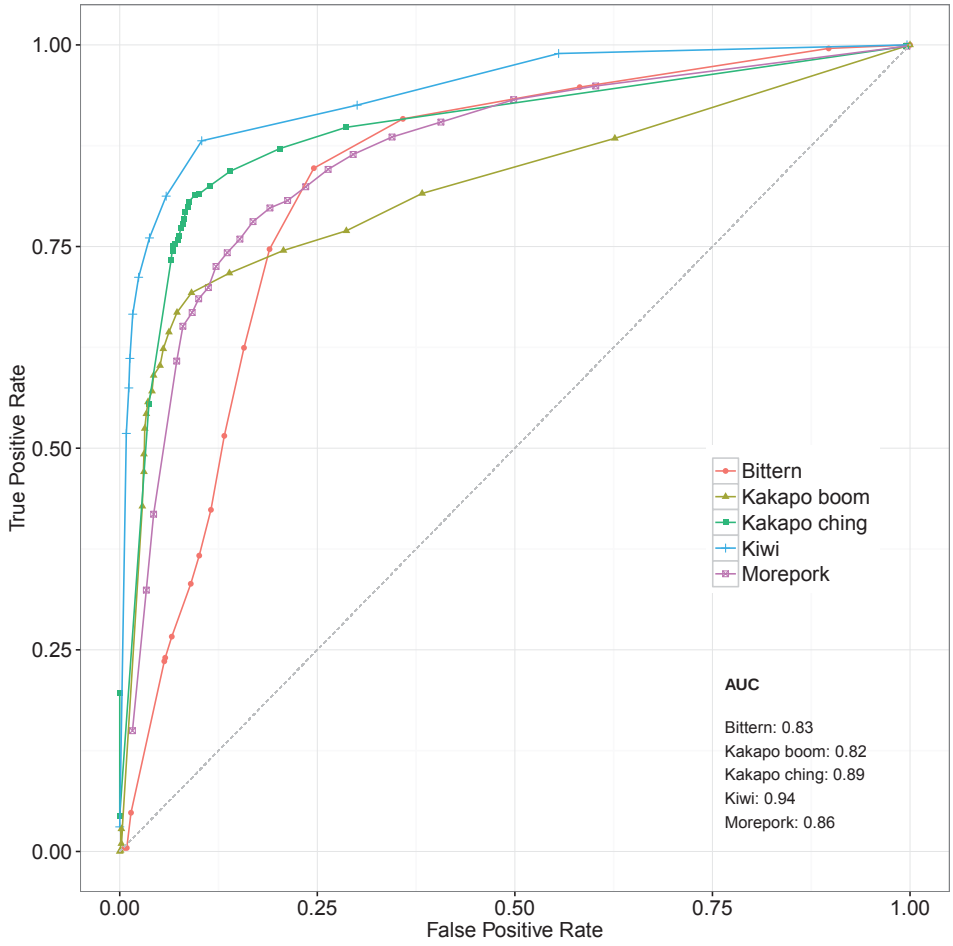


Figure 5.7: Receiver Operating Characteristic (ROC) curve of each species based on the non-experimental continuous field recordings. The straight dotted line refers to the Area Under the ROC Curve (AUC) of 0.5.

and rejecting other sections (noise), was satisfactory at 85%. Further, the detectors were capable of avoiding most of the sections where the species was not vocalising (86% specificity) which is particularly important when analysing long continuous unattended recordings.

The bittern detector was able to detect almost all the close-range booms and inhalations. The recall gradually decreased when the boom became more faded (down to 78% with 'very faded' booms). Nearly half of the extremely faded booms were detected by the algorithm. The specificity and accuracy were high (81%), but the precision was low (21%). The kākāpō and brown kiwi detectors yielded more than 75% precision. Overall, only 60% of detected sounds were target bird sounds, meaning that post processing is needed to weed out false positives, particularly for bittern and morepork (21% and 48% precision respectively).

Table 5.3: Detection results on the complete test dataset of the four species of birds used in this study. The recall depends on the quality of the bird vocalisations. TP= true positives and GT=ground truth, calculated using the number of seconds that contain the target bird sounds according to human experts.

Quality		Very loud	Loud	Faded	Very faded	Extremely faded
Bittern (1 hr)	**TP (sec)**	20	47	29	43	34
	GT (sec)	20	49	34	55	71
	Recall	100%	96%	85%	78%	48%
Kākāpō boom (30 min)	**TP (sec)**	224	179	86	42	57
	GT (sec)	224	182	164	88	162
	Recall	100%	98%	52%	48%	35%
Kākāpō ching (25 min)	**TP (sec)**	266	34	153	32	10
	GT (sec)	267	35	164	38	103
	Recall	100%	97%	93%	84%	10%
Brown kiwi (1 hr 25 min)	**TP (sec)**	139	129	129	65	5
	GT (sec)	142	137	190	113	42
	Recall	98%	94%	68%	58%	12%
Morepork (45 min)	**TP (sec)**	72	266	118	79	25
	GT (sec)	76	290	134	93	55
	Recall	95%	92%	88%	85%	45%
Total (4 hrs 5 min)	**Recall**	**99%**	**95%**	**75%**	**67%**	**30%**

5.3.3 Comparison With Other Methods

We evaluated our method with the common energy based thresholding (Jinnai et al., 2012) and median clipping (Lasseck, 2013, 2015b). Table 5.5 summaries our findings on the same dataset as presented in the previous section (Table 5.3 and Table 5.4). Consisting with previous tables we counted the number of seconds of target sounds detected instead of number of bird calls.

As we mentioned before, this is not a perfect comparison, because the reference methods are detecting any sound, while the wavelet filter only detect target bird sounds. Even though we cannot directly compare the other measures, the recall is still unbiased and can be compared. The overall recall of the proposed segmentation by wavelet filtering method (78%) was better than time domain energy thresholding (47%) and spectrogram-based median clipping (30%). Another method that could be used would be spectrogram cross-correlation (Cortopassi and Bradbury, 2000). However, this requires the manual selection of a large number of individual calls to act as templates, and degrades quickly with noise, and we therefore chose not to use it here.

Table 5.4: Detection results on the complete test dataset of four species of birds. Empty cells represent the recordings devoided of the target species.

Species	Sound file	TP (sec)	FP (sec)	TN (sec)	FN (sec)	Recall	Precision	Specificity	Accuracy
Bittern (1 hr)	*Test1*	38	82	773	7	84%	32%	90%	90%
	Test2	43	294	548	15	74%	13%	65%	66%
	Test3	27	133	727	13	68%	17%	85%	84%
	Test4	65	123	691	21	76%	35%	85%	84%
						76%	**21%**	**81%**	**81%**
Kākāpō boom (30 min)	*Test1*	97	4	196	3	97%	96%	98%	98%
	Test2	93	3	172	32	74%	97%	98%	88%
	Test3	138	16	77	69	67%	90%	83%	72%
	Test4	105	57	122	16	87%	65%	68%	76%
	Test5	30	56	214	0	100%	35%	79%	81%
	Test6	125	0	63	112	53%	100%	100%	63%
						72%	**81%**	**86%**	**80%**
Kākāpō ching (25 min)	*Test1*	105	7	164	24	81%	94%	96%	90%
	Test2	130	0	139	31	81%	100%	100%	90%
	Test3	110	0	176	14	89%	100%	100%	95%
	Test4	90	0	194	16	85%	100%	100%	95%
	Test5	60	83	129	28	68%	42%	61%	63%
						81%	**85%**	**90%**	**86%**
Brown kiwi (1 hr 25 min)	*Test1*	25	2	271	2	93%	93%	99%	99%
	Test2	0	4	296	0		0%	99%	99%
	Test3	0	62	238	0		0%	79%	79%
	Test4	60	5	203	32	65%	92%	98%	88%
	Test5	35	6	258	1	97%	85%	98%	98%
	Test6	0	5	295	0		0%	98%	98%
	Test7	69	0	200	31	69%	100%	100%	90%
	Test8	20	5	258	17	54%	80%	98%	93%
	Test9	31	4	265	0	100%	89%	99%	99%
	Test10	34	0	261	5	87%	100%	100%	98%
	Test11	41	0	256	3	93%	100%	100%	99%
	Test12	36	31	229	4	90%	54%	88%	88%
	Test13	8	16	248	28	22%	33%	94%	85%
	Test14	32	1	267	0	100%	97%	100%	100%
	Test15	33	2	235	30	52%	94%	99%	89%
	Test16	34	8	254	4	89%	81%	97%	96%
	Test17	41	16	243	0	100%	72%	94%	95%
						76%	**75%**	**96%**	**94%**
Morepork (45 min)	*Test1*	45	4	247	4	92%	92%	98%	97%
	Test2	97	9	162	32	75%	92%	95%	86%
	Test3	75	109	113	3	96%	41%	51%	63%
	Test4	41	190	57	12	77%	18%	23%	33%
	Test5	91	12	177	20	82%	88%	94%	89%
	Test6	147	70	71	12	92%	68%	50%	73%
	Test7	47	27	223	3	94%	64%	89%	90%
	Test8	4	160	135	1	80%	2%	46%	46%
	Test9	13	25	261	1	93%	34%	91%	91%
						86%	**48%**	**70%**	**74%**
Total (4 hrs 5 min)		**2,315**	**1,631**	**10,108**	**646**	**78%**	**59%**	**86%**	**85%**

Table 5.5: Recall, precision, specificity, and accuracy from our algorithm, energy thresholding, and median clipping on the complete test dataset of four species of birds.

Species	Sound file	TP (sec)	FP (sec)	TN (sec)	FN (sec)	Recall	Precision	Specificity	Accuracy
Bittern (1 hr)	This study	173	632	2,739	56	76%	21%	81%	81%
	Energy thresholding	119	746	2,625	110	52%	14%	78%	76%
	Median clipping	74	934	2,437	155	32%	7%	72%	70%
Kākāpō boom (30 min)	This study	588	136	844	232	72%	81%	86%	80%
	Energy thresholding	242	13	967	578	30%	95%	99%	67%
	Median clipping	89	62	918	731	11%	59%	94%	56%
Kākāpō ching (25 min)	This study	495	90	802	113	81%	85%	90%	86%
	Energy thresholding	309	76	817	298	51%	80%	91%	75%
	Median clipping	150	30	863	457	25%	83%	97%	68%
Brown kiwi (1hr 25 min)	This study	499	167	4,277	157	76%	75%	96%	94%
	Energy thresholding	327	2,890	1,554	329	50%	10%	35%	37%
	Median clipping	375	59	4,385	281	57%	86%	99%	93%
Morepork (45 min)	This study	560	606	1,446	88	86%	48%	70%	74%
	Energy thresholding	404	1007	1,045	244	62%	29%	51%	54%
	Median clipping	204	56	1,996	444	31%	78%	97%	81%
Total (4 hrs 5 min)	**This study**	2,315	1,631	10,108	646	**78%**	**59%**	**86%**	**85%**
	Energy thresholding	1,401	4,732	7,008	1,559	**47%**	**23%**	**60%**	**57%**
	Median clipping	892	1,141	10,599	2,068	**30%**	**44%**	**90%**	**78%**

5.3.4 Results on the RMBL Robin Database

We picked the very first recording to train the wavelet call detection algorithm (without exploring the dataset) and all other 38 recordings were treated as test data. The syllable level annotations were used to generate the binary ground truth, presence/absence of robin in each second of recoding. Wavelet filtering algorithm detected 2,479 seconds of robin songs out of 2,558 total robin seconds (97% recall). To compare with two reference methods the recall was converted into syllable level and found 99.7% recall compared to 91.3% (Potamitis et al., 2014) and to 76.0% (song level; Chu and Blumstein, 2011). The precision gained by the proposed wavelet filtering (88%) was higher than (Potamitis et al., 2014) (71%; syllable level) and (Chu and Blumstein, 2011) (75%; song level). The specificity and accuracy of our method were 87% and 92% respectively.

5.4 Discussion

Automated recordings of forest soundscape combined with computer-based acoustic analysis have the potential to provide a scalable and cost-effective method for monitoring bird populations. In order to reliably recognise bird species that are calling in long recordings, it is necessary to deal with the noise that is present in the recordings, and reliably segment the bird sounds even when they originated a significant distance from the microphone, and therefore the sound is substantially degraded. In this paper, we have described a method of filtering out calls of a target species from recordings by using a simple classifier based on the coefficients of the wavelet packet decomposition. Our intention is that the method should degrade gracefully with reducing quality of call. To this end, we manually labelled the quality of each call recorded into five classes, from close to the microphone to extremely faded.

With the help of a playback experiment we demonstrated that the proximity of the bird (actually a speaker for this experiment) to the recorder has a strong effect on recall, which was highest (95%) when the bird was facing the recorder at a short distance (20 m), and lowest (62%) when the bird was calling in the direction away from the recorder and relatively far (120m) from it. Looking at the spectrograms, only the fundamental frequency of the kiwi and morepork calls were visible at 120 m, but approximately 75% of them were successfully detected by our method.

Given that the algorithm presented here is tailored to filter the target species, meaning that some form of classification is done by the algorithm, the question would be how good is the method at avoiding other species in the background. In the playback dataset (three 5 min recordings, each was composed of four species), the precision of the brown kiwi, morepork, bittern, kākāpō boom, and kākāpō chinging wavelet filters were 61%, 69%, 78%, 7%, and 17% respectively. Misclassification of some trill sounds of morepork as kiwi was observed, but none of the more-pork sound was confused with kiwi. The chinging sound of the kākāpō was also misclassified with kiwi, but as they do not currently co-exist in the same habitat it is not expected to be a problem. When searching for morepork, the only misclassification occurred with some kiwi calls, particularly the female calls. Notably, all the false positives in bittern detection were generated by inter-syllable gaps within each booming bout due to smoothing of the energy curve generated on the wavelet coefficients (Fig. 5.6) and rich time resolution (1 sec) used. However, generally this smoothing of the energy curve (rather than thresholding the wavelet coefficients themselves) was helpful to avoid large number of false positives. According to the results, the worst case was kākāpō. When searching for kākāpō booms while some bittern booms were detected as kākāpō all the other false positives occurred due to low frequency background noise. Some morepork and brown kiwi calls were misclassified as kākāpō chinging.

Compared to the aforementioned experimental dataset, the non-experimental continuous dataset of kiwi gained better precision (75%); the false positives were mainly caused by farm animals (e.g. donkeys and roosters), and random click sounds (during heavy rains). These random events can be easily removed by simple post-processing of the results (imposing a minimum length) because kiwi produce sequence of syllables instead of isolated syllables. Most of the aeroplane, wind, and thunder were successfully removed during wavelet denoising. The precision of morepork detection was just below 50%, the false positives were due to heavy rain and thunder that affected beyond low frequencies in the spectrogram, and high level of other background noise. Optimised wavelet node lists of kiwi and morepork overlapped slightly. Some morepork calls, in particular the trill sounds were detected as kiwi and vice versa when the morepork dataset and kiwi dataset were input to kiwi and morepork detectors interchangeably. In contrast to playback results, both kākāpō detectors (boom and ching) demonstrated more reliable precision (more than 80%) in the presence of recordings collected in their natural habitats.

The false positive rate of bittern was higher (21% precision) compared to the other species and mainly caused by high level of noise of vehicles passing, aeroplanes, and human. However, 81% specificity and 76% recall of bittern detection confirmed that this method is useful to initially avoid sections devoid of bittern sounds while preserving most of the bittern sounds. The process can be followed by manual or automated classification of the detected presumed segments to improve the precision. Its also important to remind that recall and precision are inversely related to each other and we can adjust them changing the threshold under different circumstances, e.g. when estimating the population trend we can increase the precision sacrificing some faded bittern sounds or we can improve the recall at the cost of some false positives when detecting the presence/absence of bittern in a suspected area. Another option we suggest for practitioners that currently use spectrogram scanning of bittern for surveying purposes is to use the denoised recordings instead of noisy original recordings. The manual scanning process would be significantly faster, sensitive and less painful to eyes/ears if the denoised recordings are used (Fig. 5.8). It is also likely that using a better classifier than simply the logical or of the selected wavelet nodes would improve classification, and we intend to develop this further in future work.

Our intention is that the method should degrade gracefully with reducing quality of call. To this end, we manually labelled the quality of each call recorded into five classes, from close-to-microphone (very loud) to extremely faded. Our experiments revealed that our method is highly sensitive to very loud calls (99%), then decreased slightly to 67% in presence of very faded calls, and reasonably to 30% when the calls are extremely faded in the recording.

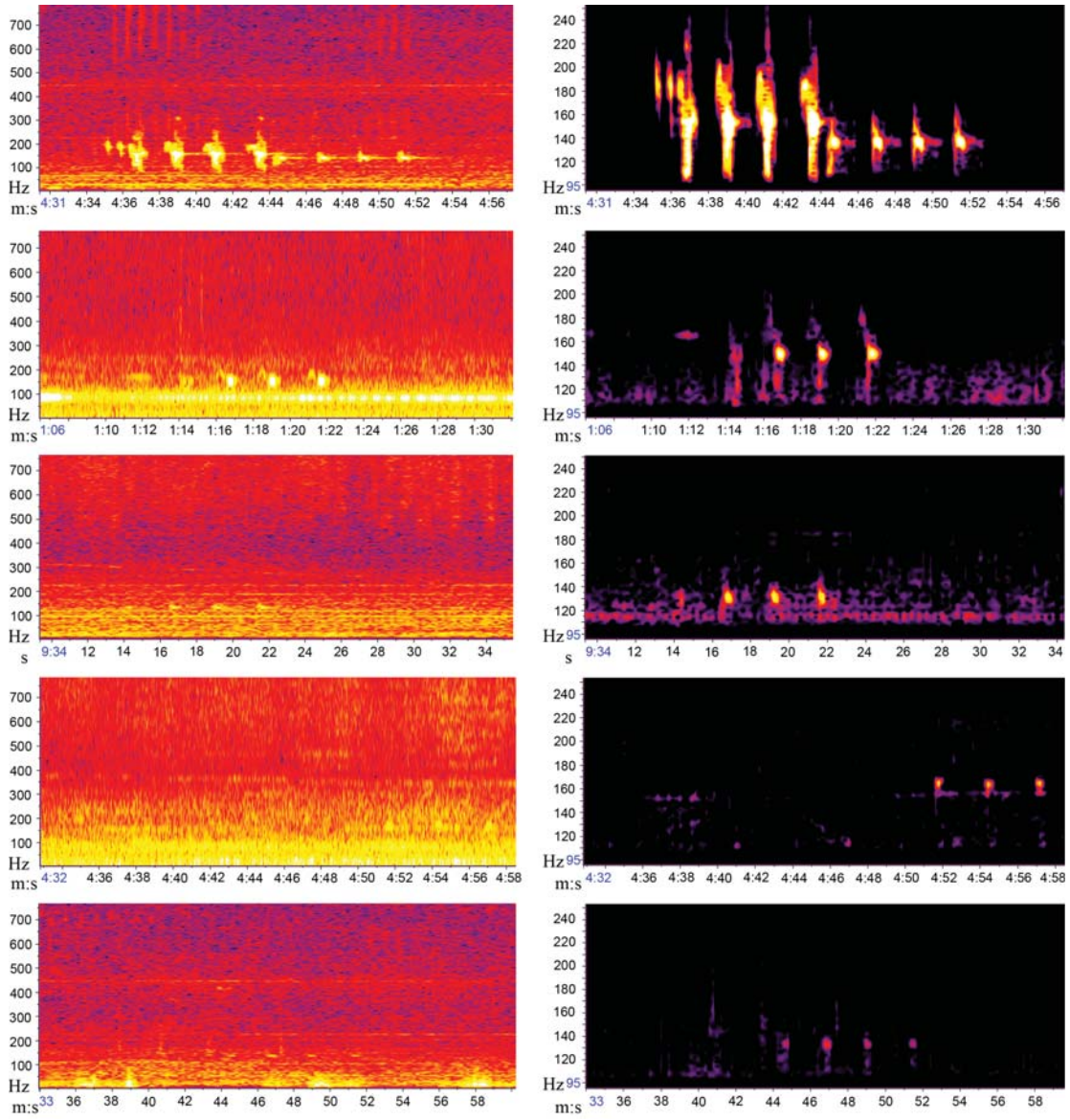


Figure 5.8: Original and denoised bittern boom examples used in this study ranging from very close (first row) to extremely faded (last row). On the left, the spectrogram settings are as given in (O'Donnell and Williams, 2015). On the right, another window preset with window size (sharpness) 400 was defined in Raven software to visualise the denoised bittern booms focusing on the bird's frequency range.

The comparison with the two commonly-used methods confirmed that the proposed wavelet call detection algorithm was significantly better in terms of recall, precision, specificity, and accuracy. Training the algorithm on a new species simply requires some manually labelled call examples. For instance, we tested the method with RMBL robin dataset and gained better results than Potamitis et al. (2014) and Chu and Blumstein (2011).

The detection algorithm took less than half a second to process a 5 min long file (approximately 20 min to process more than 4 hrs of test dataset) on a Core i7 2.4 GHz computer running Matlab code. This saves substantial time for wildlife managers and community groups to spend on other things. We intend to devise methods to detect when the sounds are faded (based on signal-to-noise computations) and offer the chance for humans to check these cases manually. However, in general we believe that this method is suitable and ready for use by conservationists.

Another advantage is that the training audios also can contain noise, meaning that natural field recordings can be used, hence the practical use of the method is high. If someone wants to train the method for a particular species, quality of the annotation demands expertise, but we believe an average person should be able to get most of the work done by carefully analysing the spectrograms. The amount of training data primarily depends on the repertoire of the particular species but also important to include different quality birdsong (e.g. very close to extremely faded). Even though, we used only one sound file (approximately 5 min long) to train the method to detect almost all (99.7%) robin in the RMBL dataset. Favourably, this annotation needs to be done only once; after training the method on those example recordings, one can process field recordings of any size to find the target species. Of course, it is necessary to evaluate the trained algorithm on a set of field recordings before processing the whole dataset to decide whether the training data are adequate or more example recordings are needed.

The development of precisely annotated datasets and making them available to other researchers to compare their methods with is essential to improve this field of research (Chapter 2). Therefore, we make available the dataset used in this study with their annotations (S 5.3–S 5.6 Audio and annotation) from `http://avianz.massey.ac.nz` (and also from GitHub `https://github.com/smarsland/AviaNZ`).

5.5 Supporting Information

S 5.1 Audio. Birdsong examples in Fig. 5.4

S 5.2 Audio. Australasian bittern example in Fig. 5.6

S 5.3 Audio and annotation. Australasian bittern dataset

S 5.4 Audio and annotation. Kākāpō dataset

S 5.5 Audio and annotation. Brown kiwi dataset

S 5.6 Audio and annotation. Morepork dataset

Acknowledgments

We appreciate the support given by Emma Williams throughout the study providing a large number of bittern sound files already labelled. Authors would like to thank Bruce C. Robertson, Andrew Digby and the kākāpō recovery team at Department of Conservation for providing their recordings, and Alex Brighten for morepork and kiwi recordings. This research was partially funded by HETC, Sri Lanka (KLN/O-Sci/N6), School of Engineering and Advanced Technology, Massey University, New Zealand, and NP was awarded with the runner up prize (CIA 14/05) in WWF Conservation Innovation Awards 2014 New Ideas for Nature (research innovation). Recordings by IC were made with funding from Massey University Research Fund, 2011 (RM1000015982 P-MURF, 7003-Massey University). NP was supported by University of Kelaniya, Sri Lanka.

Chapter 6

Concluding Remarks

During the last decade, conservation managers have embraced passive acoustic monitoring as a way to monitor wildlife on a large-scale. Recordings are routinely collected from key sites (e.g. remote areas or ecologically sensitive areas) to monitor species; the resulting data have the potential to enhance management: detecting the presence of species, and potentially estimating the size of populations based on the amount of acoustic activity recorded to systematically observe the trend of population sizes over time. The greatest bottleneck in this process is the need for very costly manual scanning of large amounts of continuous field recordings. To date, there is a lack of robust automated methods to screen these recordings. This thesis aimed to produce methods to automatically recognise target bird species from field recordings collected with autonomous recorders.

It quickly became clear that the state of the art methods were not mature enough to perform birdsong recognition to the accuracy that is required by the conservationists on the kind of field recordings collected by passive acoustic recording devices. The primary problem was noise: the literature review revealed that more than two thirds of the work done so far has limited their scope to analyse less noisy and carefully selected recordings, and is not ready to go beyond laboratory tests or to process naturally noisy continuous field recordings (Chapter 2). Conjointly with noise and the (lack of) proximity of the birds to recorders in their natural habitats, detecting bird sounds from continuous recordings suffered not only from low recall but also from a high false positive rate.

6.1 Pre-processing of Field Recordings

Typical noise in field recordings includes wind, rain, interference of competing species, insects, human speech, aeroplane, and other machine noises. While the band-pass filter can remove noise that does not overlap with the frequency range of the target species

or the type of song of interest, band-pass filters are unable to treat noise that mixes with target sounds. An alternative based on wavelet denoising was devised in Chapter 4, and found to be able to remove any quasi-stationary noise overlapped with birdsong regardless of the frequency of interest. Testing noise reduction methods requires the use of real data, not the manual corruption of clean recordings, because natural noise is not mixed linearly with the signal. On such data, the results in Chapter 4 confirmed an order of magnitude improvement in the signal-to-noise ratio.

While the denoising method was devised for automatic recorders, it can be used for any type of recording, and could potentially help as a pre-processing step for any acoustic recognition task. A logical next step would be to investigate how the denoising affected the recognition accuracy of birdsong recognition, whether manual or automatic; this is left for future work.

6.2 Segmentation and Recognition

Generally, automated birdsong recognition follows a standard approach (in common with other areas of acoustic analysis, such as human speech recognition) that begins with noise treatment and proceeds with segmentation of the vocalisations from the recordings. When analysing field recordings segmentation is very important because in general birds are audible for only a minority of the recorded time and much of the recording is effectively silent. Following segmentation, features are derived from the filtered segments and are input to a classifier which performs the actual recognition.

In Chapter 5 a novel method of simultaneously segmenting and recognising the song of a particular target species was developed. Like the noise reduction method, it is based on wavelets, which provide a useful representation of sounds that is robust to noise. The method was trained on the calls of several target bird species, based on standard field recordings. This segmentation by wavelet filtering method was able to separate the target bird sounds from the non-stationary noise (that was left after denoising) with greater than 80% AUC.

While the method is species-specific, it is easy to apply it to a new species: all that is required is a training set demonstrating exemplars of the calls of the target bird. In contrast, many of the methods in the literature depend upon specific features of the call of the target species, and are thus hard to retrain for other species.

6.3 Protocols for Data Acquisition

Automated recorders are routinely available, hence collecting acoustic data is now reasonably easy. However, compared to the ease of collecting the data, huge effort is required to analyse those recordings and to interpret the findings in order to ultimately

estimate the size of the bird populations. Even though scheduling the recorders and mounting them in the field seems to be easy, the current practice is not robust. Different people use different methods based on their knowledge and their own preferences. Based on the evidence in Chapter 3, which investigated the environmental factors that affect birdsong acquisition using omini-directional automated recorders, variations in the method will cause changes in the recordings, and hence the interpretation of the results. There is a lack of protocols for data collection that directly effects the decisions made based on this data. Therefore, it is critical that standard protocols for bioacoustic data acquisition using autonomous recorders are developed. These protocols need to be compatible with the behaviour and ecology of the birds, their distribution and habitat, and the nature of the vocalisations – hence species-specific. While there are protocols established for few species, in particular the endangered species, for example Australasian bittern (*Botaurus poiciloptilus*) (O'Donnell and Williams, 2015), there are many more species to be considered. The key questions in data collection are when to record, where to deploy the recorders, how many recorders are needed to cover a unit area, the distance between the recorders, and the height of the recorder from the ground. The scenario is complex and the answers will heavily depend on the species being monitored and the geographical location of the study area as well as the current knowledge of the population sizes. There is also the need for more field-based experimental work to fully understand the best positioning of the recorders to collect the most high-quality with minimal recorders.

6.4 Making Birdsong Recognition Practical: Recommendations

As a final contribution to the area of birdsong recognition, we consider briefly a few key points that should be considered by researchers in order to improve the comparison methods of new developments, and to make them more useful to conservation practitioners.

6.4.1 Shared Data

As has been clearly shown in this thesis, the quality of the acoustic data is critical for automated birdsong recognition: the chance of failure is very high if a method is assessed only on good quality recordings before testing in the field. In addition, it is hard to compare methods when different types and quality of data are used. While the datasets for the various competitions like BirdCLEF can help with this, they are not currently based on long, noisy field recordings. Therefore we emphasise the need of departing from 'private data' in order to develop this field of research to the extent

that is required by wildlife managers and ecologists.

In addition to the raw recordings it is essential to accompany them with their human (ground truth) annotations, which requires some level of expertise to make them correct. Most of the spectrogram analysis software (e.g. Raven and Praat) provide the facility of creating annotation tables. The other option would be to record the time stamps (start and the end of each birdsong unit) into a spreadsheet.

6.4.2 Inventory Management of Surveys

The need for handling big data becomes a problem when passive acoustic monitoring is employed at large-scale over the long-term and recordings quickly accumulate from multiple recorders at multiple sites. While choosing the optimum sampling rate and bit rate can avoid generating large datasets unnecessarily, acoustic data are still significant in size.

Consistency over time is crucial in passive acoustic monitoring, particularly when surveys are done to observe the population trend over time. There are surveys carried out to assess the efficacy of conservation efforts (e.g. before and after a pest control season) using passive acoustic monitoring of birds/a population of bird.

The transition from the current manual spectrogram reading to the automated processing of recordings using a software tool would save huge amount of human time while improving the consistency of results. However, we do not recommend the direct transition from manual to automatic analysis. For example, consider the case where a sound file is analysed now using currently-available software, and again in 10 years, after a lot of further research. If the number of calls detected in the same file were compared, presumably they would be quite different as more accurate methods will be devised. Therefore, using the calls detected now to provide an abundance index could be risky without careful manual analysis and comparison with the machine accuracy. For example, randomly validating the results generated by the software would be useful to ensure the software is doing its job smoothly.

6.4.3 Community Participation in Conservation

Local communities enjoy bird watching and are interested in being involved in bird conservation. Community groups are a great resource for a country. In New Zealand, where large amounts of predator-based work is necessary to re-establish the avifauna, this is particularly clear. Currently, community groups play the largest role in pest control in New Zealand after the Department of Conservation.

In most cases, these groups are driven by their own dedication and interest. However, the lack of subject knowledge hinders their performance. For example, according to our experience, most of them are not aware of how to process (e.g. spectrogram

reading) the recordings collected to monitor the consequences of their efforts in pest control. Deploying the recorders is also a puzzle for them. Therefore, they largely need support from experts and more experienced users. For instance, providing them training opportunities and extending the underline communication with other groups and the Department of Conservation is useful. Community involvement and awareness would be a key to achieve large-scale goals such as *Predator Free New Zealand 2050*.

Acoustic monitoring of birds has the potential to provide a lot of useful data to ecologists, wildlife managers, and anybody else with an interest in the state of our birdlife. This thesis has made original contributions to this. However, there is still significant work to be done before automatic passive acoustic monitoring of our birds is a solved problem. This thesis has provided a foundation for this future challenge.

Appendix A

Generalised Linear Models

A.1 Analysis I - Model effects

Table A.1: Model effects. Dependent variable is SnNR.

Call example

bf
(brown kiwi female)

Tests of Model Effects - bf

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	435926.413	1	.000
DayNight	458.940	1	.000
OpenForest	1227.892	1	.000
Height	455.003	1	.000
RDirection	62.909	3	.000
Distance	5336.644	4	.000
DayNight * OpenForest	12.557	1	.000
DayNight * Height	37.118	1	.000
DayNight * RDirection	25.307	3	.000
DayNight * Distance	35.076	4	.000
OpenForest * Height	10.088	1	.001
OpenForest * RDirection	31.136	3	.000
OpenForest * Distance	618.507	4	.000
Height * Distance	23.324	4	.000
RDirection * Distance	94.525	12	.000

bm1
(brown kiwi male)

Tests of Model Effects - bm1

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	391762.615	1	.000
DayNight	8.414	1	.004
OpenForest	276.372	1	.000
RDirection	39.304	3	.000
Distance	1545.810	4	.000
DayNight * RDirection	30.015	3	.000
DayNight * Distance	105.438	4	.000
OpenForest * Distance	287.972	4	.000
RDirection * Distance	73.657	12	.000

bm2
(brown kiwi male)

Tests of Model Effects - bm2

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	420063.931	1	.000
DayNight	23.049	1	.000
OpenForest	419.823	1	.000
Height	8.193	1	.004
RDirection	34.103	3	.000
Distance	2292.386	4	.000
DayNight * OpenForest	11.669	1	.001
DayNight * Height	22.944	1	.000
DayNight * RDirection	37.997	3	.000
DayNight * Distance	131.733	4	.000
OpenForest * Height	13.147	1	.000
OpenForest * Distance	418.248	4	.000
Height * Distance	27.620	4	.000
RDirection * Distance	76.511	12	.000

lskf
(little spotted kiwi female)

Tests of Model Effects - lskf

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	388443.209	1	.000
DayNight	32.719	1	.000
OpenForest	505.603	1	.000
Height	8.854	1	.003
RDirection	20.074	3	.000
Distance	1841.466	4	.000
DayNight * Height	45.319	1	.000
DayNight * RDirection	27.660	3	.000
DayNight * Distance	84.729	4	.000
OpenForest * Height	19.749	1	.000
OpenForest * Distance	276.773	4	.000
Height * Distance	47.699	4	.000
RDirection * Distance	58.597	12	.000

lskm1
(little spotted kiwi male)

Tests of Model Effects - lskm1

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	410010.457	1	.000
DayNight	54.633	1	.000
OpenForest	346.962	1	.000
RDirection	38.052	3	.000
Distance	2405.542	4	.000
DayNight * OpenForest	11.260	1	.001
DayNight * RDirection	58.313	3	.000
DayNight * Distance	143.316	4	.000
OpenForest * Distance	286.642	4	.000
RDirection * Distance	65.743	12	.000

lskm2
(little spotted kiwi male)

Tests of Model Effects - lskm2

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	368738.967	1	.000
DayNight	7.099	1	.008
OpenForest	232.039	1	.000
RDirection	21.583	3	.000
Distance	1467.397	4	.000
DayNight * OpenForest	7.374	1	.007
DayNight * RDirection	33.928	3	.000
DayNight * Distance	120.392	4	.000
OpenForest * Distance	219.223	4	.000
RDirection * Distance	73.096	12	.000

mp
(more-pork sound
of morepork)

Tests of Model Effects - mp

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	502517.002	1	.000
DayNight	54.713	1	.000
OpenForest	1295.983	1	.000
Height	131.693	1	.000
RDirection	35.373	3	.000
Distance	2681.364	4	.000
DayNight * Height	91.434	1	.000
DayNight * RDirection	37.279	3	.000
DayNight * Distance	110.050	4	.000
OpenForest * Height	12.285	1	.000
OpenForest * RDirection	21.292	3	.000
OpenForest * Distance	479.029	4	.000
Height * RDirection	13.267	3	.004
Height * Distance	154.768	4	.000
RDirection * Distance	78.997	12	.000

trilH (trill sound of morepork)

Tests of Model Effects - trilH

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	376264.971	1	.000
DayNight	15.220	1	.000
OpenForest	361.261	1	.000
Height	20.200	1	.000
Distance	1520.474	4	.000
DayNight * Height	24.568	1	.000
DayNight * Distance	99.755	4	.000
OpenForest * Height	36.902	1	.000
OpenForest * Distance	324.236	4	.000
Height * Distance	25.857	4	.000

trilL(trill sound of morepork)

Tests of Model Effects - trilL

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	351003.071	1	.000
OpenForest	193.304	1	.000
Distance	558.507	4	.000
OpenForest * Distance	146.340	4	.000

bittern

Tests of Model Effects - bittern

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	402875.183	1	.000
DayNight	65.031	1	.000
RDirection	20.533	3	.000
Distance	216.849	4	.000
DayNight * Distance	35.465	4	.000
RDirection * Distance	47.662	12	.000

kBoom (kakapo boom)

Tests of Model Effects - kBoom

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	381696.379	1	.000
DayNight	69.276	1	.000
RDirection	17.212	3	.001
BDirection	20.533	3	.000
Distance	142.734	4	.000
DayNight * BDirection	68.076	3	.000
DayNight * Distance	27.683	4	.000
RDirection * Distance	35.222	12	.000

kc (kakapo chinging)

Tests of Model Effects - kc

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	322419.016	1	.000
OpenForest	121.512	1	.000
Height	29.193	1	.000
Distance	473.442	4	.000
OpenForest * Height	24.233	1	.000
OpenForest * Distance	73.503	4	.000
Height * Distance	16.704	4	.002

weka

Tests of Model Effects - weka

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	418456.680	1	.000
DayNight	101.130	1	.000
OpenForest	587.728	1	.000
Height	69.573	1	.000
RDirection	65.460	3	.000
Distance	3064.944	4	.000
DayNight * OpenForest	22.086	1	.000
DayNight * Height	10.504	1	.001
DayNight * RDirection	21.549	3	.000
DayNight * Distance	119.356	4	.000
OpenForest * RDirection	18.605	3	.000
OpenForest * Distance	548.977	4	.000
Height * Distance	16.513	4	.002
RDirection * Distance	82.578	12	.000

kaka

Tests of Model Effects - kaka

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	358669.103	1	.000
DayNight	9.937	1	.002
OpenForest	358.353	1	.000
Distance	1063.572	4	.000
DayNight * Distance	87.049	4	.000
OpenForest * Distance	183.242	4	.000

hihi

Tests of Model Effects - hihi

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	270657.482	1	.000
OpenForest	99.577	1	.000
Height	30.223	1	.000
Distance	374.643	4	.000
OpenForest * Height	6.496	1	.011
OpenForest * Distance	54.688	4	.000

robin

Tests of Model Effects - robin

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	261386.084	1	.000
OpenForest	91.663	1	.000
Height	7.682	1	.006
Distance	645.318	4	.000
OpenForest * Height	15.379	1	.000
OpenForest * Distance	118.373	4	.000

tui

Tests of Model Effects - tui

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	385436.967	1	.000
OpenForest	252.596	1	.000
RDirection	19.807	3	.000
Distance	923.333	4	.000
OpenForest * RDirection	19.526	3	.000
OpenForest * Distance	254.815	4	.000
RDirection * Distance	57.663	12	.000

sad1 (saddleback)

Tests of Model Effects - sad1

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	468173.099	1	.000
DayNight	15.806	1	.000
OpenForest	178.714	1	.000
RDirection	20.618	3	.000
BDirection	27.320	3	.000
Distance	1387.403	4	.000
DayNight * OpenForest	6.450	1	.011
DayNight * RDirection	43.623	3	.000
DayNight * BDirection	21.213	3	.000
DayNight * Distance	123.319	4	.000
OpenForest * BDirection	25.179	3	.000
OpenForest * Distance	240.741	4	.000
RDirection * BDirection	155.886	9	.000
RDirection * Distance	87.001	12	.000

sad2 (saddleback)

Tests of Model Effects - sad2

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	386868.562	1	.000
OpenForest	122.999	1	.000
RDirection	29.193	3	.000
BDirection	23.945	3	.000
Distance	916.648	4	.000
OpenForest * Distance	176.392	4	.000
RDirection * BDirection	179.329	9	.000
RDirection * Distance	70.407	12	.000

sad3 (saddleback)

Tests of Model Effects - sad3

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	463032.976	1	.000
OpenForest	116.773	1	.000
RDirection	47.418	3	.000
BDirection	34.751	3	.000
Distance	1328.026	4	.000
OpenForest * RDirection	12.942	3	.005
OpenForest * BDirection	27.592	3	.000
OpenForest * Distance	301.649	4	.000
RDirection * BDirection	126.869	9	.000
RDirection * Distance	71.048	12	.000

Table A.2: EMM – Open vs Forest – overall test results – significant effects after sequential Sidak correction on $\alpha = 0.01$.

Call example	Wald Chi-Square	df	Significance
bf	1262.722	1	0.000
bm1	275.152	1	0.000
bm2	424.367	1	0.000
lskf	504.704	1	0.000
lskm1	344.597	1	0.000
lskm2	231.155	1	0.000
mp	1274.189	1	0.000
trilH	363.747	1	0.000
trilL	190.194	1	0.000
bittern	Experiment site was not in the model		
kBoom	Experiment site was not in the model		
kc	119.971	1	0.000
weka	598.767	1	0.000
kaka	350.096	1	0.000
hihi	98.226	1	0.000
robin	91.299	1	0.000
tui	251.081	1	0.000
sad1	177.674	1	0.000
sad2	121.402	1	0.000
sad3	115.682	1	0.000

Table A.3: EMM – Day vs Night – overall test results – significant effects after sequential Sidak correction on $\alpha = 0.01$.

Call example	Wald Chi-Square	df	Significance
bf	465.943	1	0.000
bm1	8.441	1	0.004
bm2	23.136	1	0.000
lskf	32.878	1	0.000
lskm1	55.147	1	0.000
lskm2	7.125	1	0.008
mp	55.008	1	0.000
trilH	15.282	1	0.000
trilL	Time of day was not in the model		
bittern	63.653	1	0.000
kBoom	68.178	1	0.000
kc	Time of day was not in the model		
weka	101.695	1	0.000
kaka	9.934	1	0.002
hihi	Time of day was not in the model		
robin	Time of day was not in the model		
tui	Time of day was not in the model		
sad1	15.886	1	0.000
sad2	Time of day was not in the model		
sad3	Time of day was not in the model		

Table A.4: EMM – Low vs High transmission height – overall test results – significant effects after sequential Sidak correction on $\alpha = 0.01$.

Call example	Wald Chi-Square	df	Significance
bf	450.625	1	0.000
bm1	Transmission height was not in the model		
bm2	8.207	1	0.004
lskf	8.861	1	0.003
lskm1	Transmission height was not in the model		
lskm2	Transmission height was not in the model		
mp	133.562	1	0.000
trilH	20.228	1	0.000
trilL	Transmission height was not in the model		
bittern	Transmission height was not in the model		
kBoom	Transmission height was not in the model		
kc	29.041	1	0.000
weka	69.261	1	0.000
kaka	Transmission height was not in the model		
hihi	29.929	1	0.000
robin	7.673	1	0.006
tui	Transmission height was not in the model		
sad1	Transmission height was not in the model		
sad2	Transmission height was not in the model		
sad3	Transmission height was not in the model		

Table A.5: EMM – Distance (20m, 25m, 50m, 100m, 120m) – individual test results – significant effects after sequential Sidak correction on $\alpha = 0.01$.

Call example	Individual Test Results	Contrast Estimate	Std. Error	Wald Chi-Square	df	Significance
bf	25m vs. 20m	-0.055	0.007	71.325	1	0.000
	50m vs. 25m	-0.159	0.005	896.085	1	0.000
	100m vs. 50m	-0.225	0.004	2522.045	1	0.000
	120m vs. 100m	-0.183	0.004	2300.891	1	0.000
bm1	25m vs. 20m	-0.045	0.006	55.449	1	0.000
	50m vs. 25m	-0.109	0.005	547.44	1	0.000
	100m vs. 50m	-0.111	0.004	746.607	1	0.000
	120m vs. 100m	-0.077	0.004	353.357	1	0.000
bm2	25m vs. 20m	-0.042	0.006	46.193	1	0.000
	50m vs. 25m	-0.122	0.005	643.257	1	0.000
	100m vs. 50m	-0.137	0.004	1136.891	1	0.000
	120m vs. 100m	-0.098	0.004	718.888	1	0.000
lskf	25m vs. 20m	-0.035	0.006	32.388	1	0.000
	50m vs. 25m	-0.114	0.005	573.856	1	0.000
	100m vs. 50m	-0.127	0.004	930.363	1	0.000
	120m vs. 100m	-0.093	0.004	514.513	1	0.000
lskm1	25m vs. 20m	-0.039	0.006	46.925	1	0.000
	50m vs. 25m	-0.127	0.005	725.696	1	0.000
	100m vs. 50m	-0.136	0.004	1097.447	1	0.000
	120m vs. 100m	-0.099	0.004	616.232	1	0.000
lskm2	25m vs. 20m	-0.036	0.006	35.918	1	0.000
	50m vs. 25m	-0.107	0.005	450.484	1	0.000
	100m vs. 50m	-0.11	0.004	683.962	1	0.000
	120m vs. 100m	-0.076	0.004	370.313	1	0.000
mp	25m vs. 20m	-0.036	0.005	53.744	1	0.000
	50m vs. 25m	-0.109	0.004	652.662	1	0.000
	100m vs. 50m	-0.139	0.004	1230.545	1	0.000
	120m vs. 100m	-0.104	0.004	827.859	1	0.000
trilH	25m vs. 20m	-0.043	0.006	44.388	1	0.000
	50m vs. 25m	-0.108	0.005	480.68	1	0.000
	100m vs. 50m	-0.117	0.004	789.155	1	0.000
	120m vs. 100m	-0.08	0.004	409.618	1	0.000
trilL	25m vs. 20m	-0.034	0.006	30.555	1	0.000
	50m vs. 25m	-0.068	0.005	196.043	1	0.000
	100m vs. 50m	-0.071	0.004	283.129	1	0.000
	120m vs. 100m	-0.043	0.004	106.882	1	0.000
bittern	25m vs. 20m	-0.019	0.005	14.924	1	0.000
	50m vs. 25m	-0.033	0.005	49.011	1	0.000
	100m vs. 50m	-0.033	0.004	59.773	1	0.000
	120m vs. 100m	-0.031	0.004	73.72	1	0.000
kBoom	25m vs. 20m	-0.015	0.005	10.358	1	0.001
	50m vs. 25m	-0.021	0.005	19.516	1	0.000
	100m vs. 50m	-0.03	0.005	44.455	1	0.000
	120m vs. 100m	-0.027	0.004	51.843	1	0.000

Call example	Individual Test Results	Contrast Estimate	Std. Error	Wald Chi-Square	df	Significance
kc	25m vs. 20m	-0.029	0.006	21.804	1	0.000
	50m vs. 25m	-0.068	0.005	163.922	1	0.000
	100m vs. 50m	-0.062	0.004	202.343	1	0.000
	120m vs. 100m	-0.04	0.004	90.54	1	0.000
weka	25m vs. 20m	-0.048	0.006	56.396	1	0.000
	50m vs. 25m	-0.143	0.005	933.082	1	0.000
	100m vs. 50m	-0.165	0.004	1525.535	1	0.000
	120m vs. 100m	-0.126	0.004	1044.493	1	0.000
kaka	25m vs. 20m	-0.031	0.006	26.36	1	0.000
	50m vs. 25m	-0.091	0.005	363.913	1	0.000
	100m vs. 50m	-0.097	0.004	492.763	1	0.000
	120m vs. 100m	-0.067	0.004	242.535	1	0.000
hihi	25m vs. 20m	-0.028	0.007	15.518	1	0.000
	50m vs. 25m	-0.07	0.006	151.856	1	0.000
	100m vs. 50m	-0.062	0.005	163.31	1	0.000
	120m vs. 100m	-0.039	0.005	72.942	1	0.000
robin	25m vs. 20m	-0.032	0.008	15.717	1	0.000
	50m vs. 25m	-0.098	0.006	271.591	1	0.000
	100m vs. 50m	-0.083	0.005	308.766	1	0.000
	120m vs. 100m	-0.057	0.004	160.325	1	0.000
tui	25m vs. 20m	-0.03	0.006	23.896	1	0.000
	50m vs. 25m	-0.087	0.005	357.638	1	0.000
	100m vs. 50m	-0.089	0.004	452.081	1	0.000
	120m vs. 100m	-0.059	0.004	226.307	1	0.000
sad1	25m vs. 20m	-0.037	0.005	57.091	1	0.000
	50m vs. 25m	-0.099	0.004	520.542	1	0.000
	100m vs. 50m	-0.09	0.004	522.144	1	0.000
	120m vs. 100m	-0.057	0.004	255.396	1	0.000
sad2	25m vs. 20m	-0.032	0.006	33.856	1	0.000
	50m vs. 25m	-0.091	0.005	366.279	1	0.000
	100m vs. 50m	-0.079	0.004	353.125	1	0.000
	120m vs. 100m	-0.049	0.004	158.389	1	0.000
sad3	25m vs. 20m	-0.036	0.005	49.12	1	0.000
	50m vs. 25m	-0.098	0.004	569.688	1	0.000
	100m vs. 50m	-0.089	0.004	538.632	1	0.000
	120m vs. 100m	-0.059	0.004	242.381	1	0.000

Table A.6: EMM – Distance (20m, 25m, 50m, 100m, 120m) – overall test results – significant effects after sequential Sidak correction on $\alpha = 0.01$.

Call example	Wald Chi-Square	df	Significance
bf	5120.568	4	0.000
bm1	1444.139	4	0.000
bm2	2115.337	4	0.000
lskf	1738.588	4	0.000
lskm1	2317.901	4	0.000
lskm2	1384.439	4	0.000
mp	2674.829	4	0.000
trilH	1405.971	4	0.000
trilL	530.703	4	0.000
bittern	213.906	4	0.000
kBoom	142.493	4	0.000
kc	452.054	4	0.000
weka	2848.775	4	0.000
kaka	1021.352	4	0.000
hihi	355.668	4	0.000
robin	587.483	4	0.000
tui	868.425	4	0.000
sad1	1342.053	4	0.000
sad2	878.114	4	0.000
sad3	1267.966	4	0.000

Table A.7: EMM – Open/Forest*Day/Night interaction – overall test results – significant effects after sequential Sidak correction on $\alpha = 0.01$.

Call example	Wald Chi-Square	df	Significance
bf	1930.763	3	0.000
bm1	Open/Forest*Day/Night interaction was not in the model		
bm2	504.521	3	0.000
lskf	Open/Forest*Day/Night interaction was not in the model		
lskm1	438.395	3	0.000
lskm2	247.167	3	0.000
mp	Open/Forest*Day/Night interaction was not in the model		
trilH	Open/Forest*Day/Night interaction was not in the model		
trilL	Open/Forest*Day/Night interaction was not in the model		
bittern	Open/Forest*Day/Night interaction was not in the model		
kBoom	Open/Forest*Day/Night interaction was not in the model		
kc	Open/Forest*Day/Night interaction was not in the model		
weka	186.689	3	0.000
kaka	Open/Forest*Day/Night interaction was not in the model		
hihi	Open/Forest*Day/Night interaction was not in the model		
robin	Open/Forest*Day/Night interaction was not in the model		
tui	Open/Forest*Day/Night interaction was not in the model		
sad1	214.003	3	0.000
sad2	Open/Forest*Day/Night interaction was not in the model		
sad3	Open/Forest*Day/Night interaction was not in the model		

Table A.8: EMM – Open/Forest*Low/High transmission interaction – overall test results – significant effects after sequential Sidak correction on $\alpha = 0.01$.

Call example	Wald Chi-Square	df	Significance
bf	1603.428	3	0.000
bm1	Open/Forest*Low/High was not in the model		
bm2	460.380	3	0.000
lskf	531.270	3	0.000
lskm1	Open/Forest*Low/High was not in the model		
lskm2	Open/Forest*Low/High was not in the model		
mp	1558.386	3	0.000
trilH	420.936	3	0.000
trilL	Open/Forest*Low/High was not in the model		
bittern	Open/Forest*Low/High was not in the model		
kBoom	Open/Forest*Low/High was not in the model		
kc	212.781	3	0.000
weka	Open/Forest*Low/High was not in the model		
kaka	Open/Forest*Low/High was not in the model		
hihi	159.311	3	0.000
robin	127.642	3	0.000
tui	Open/Forest*Low/High was not in the model		
sad1	Open/Forest*Low/High was not in the model		
sad2	Open/Forest*Low/High was not in the model		
sad3	Open/Forest*Low/High was not in the model		

Table A.9: EMM – Day/Night*Low/High transmission interaction – overall test results – significant effects after sequential Sidak correction on $\alpha = 0.01$.

Call example	Wald Chi-Square	df	Significance
bf	956.009	3	0.000
bm1	Day/Night*Low/High was not in the model		
bm2	61.534	3	0.000
lskf	92.055	3	0.000
lskm1	Day/Night*Low/High was not in the model		
lskm2	Day/Night*Low/High was not in the model		
mp	412.639	3	0.000
trilH	65.953	3	0.000
trilL	Day/Night*Low/High was not in the model		
bittern	Day/Night*Low/High was not in the model		
kBoom	Day/Night*Low/High was not in the model		
kc	Day/Night*Low/High was not in the model		
weka	186.689	3	0.000
kaka	Day/Night*Low/High was not in the model		
hihi	Day/Night*Low/High was not in the model		
robin	Day/Night*Low/High was not in the model		
tui	Day/Night*Low/High was not in the model		
sad1	Day/Night*Low/High was not in the model		
sad2	Day/Night*Low/High was not in the model		
sad3	Day/Night*Low/High was not in the model		

Table A.10: EMM – Open/Forest*Distance interaction – overall test results – significant effects after sequential Sidak correction on $\alpha = 0.01$.

Call example	Wald Chi-Square	df	Significance
bf	8550.861	9	0.000
bm1	2371.480	9	0.000
bm2	4058.628	9	0.000
lskf	3144.896	9	0.000
lskm1	3754.522	9	0.000
lskm2	2295.918	9	0.000
mp	6358.492	9	0.000
trilH	2486.788	9	0.000
trilL	760.840	9	0.000
bittern	Open/Forest*Distance was not in the model		
kBoom	Open/Forest*Distance was not in the model		
kc	Open/Forest*Distance was not in the model		
weka	5028.876	9	0.000
kaka	1723.986	9	0.000
hihi	564.019	9	0.000
robin	896.109	9	0.000
tui	1292.428	9	0.000
sad1	2344.956	9	0.000
sad2	1304.824	9	0.000
sad3	1632.038	9	0.000

A.2 Analysis II

Table A.11: EMM (Analysis II) – Recorder direction – overall test results – significant effects after sequential Sidak correction on $\alpha = 0.01$

Call example	Wald Chi-Square	df	Significance
bf	14.200	3	0.003
bm1	24.095	3	0.000
bm2	15.940	3	0.001
lskf	21.661	3	0.000
lskm1	17.499	3	0.001
lskm2	20.604	3	0.000
mp	7.138	3	0.068
trilH	18.879	3	0.000
trilL	12.643	3	0.005
bittern	11.899	3	0.008
kBoom	5.538	3	0.136
kc	24.922	3	0.000
weka	18.863	3	0.000
kaka	21.825	3	0.000
hihi	28.167	3	0.000
robin	29.154	3	0.000
tui	19.779	3	0.000
sad1	20.599	3	0.000
sad2	30.529	3	0.000
sad3	19.396	3	0.000

A.3 Analysis III

Table A.12: EMM (Analysis III) – Recorder direction – overall test results – significant effects after sequential Sidak correction on $\alpha = 0.01$

Call example	Wald Chi-Square	df	Significance
bf	23.979	3	0.000
bm1	24.474	3	0.000
bm2	17.195	3	0.001
lskf	13.472	3	0.004
lskm1	14.470	3	0.002
lskm2	15.239	3	0.002
mp	32.559	3	0.000
trilH	16.437	3	0.001
trilL	20.555	3	0.000
bittern	23.598	3	0.000
kBoom	7.587	3	0.055
kc	20.016	3	0.000
weka	20.853	3	0.000
kaka	37.682	3	0.000
hihi	32.495	3	0.000
robin	16.779	3	0.001
tui	30.388	3	0.000
sad1	17.507	3	0.001
sad2	63.109	3	0.000
sad3	24.694	3	0.000

Table A.13: EMM (Analysis III) – Wind level – individual test results – significant effects after sequential Sidak correction on $\alpha = 0.01$

Call example	Individual test results	Contrast Estimate	Std. Error	Wald Chi-Square	df	Significance
bf	moderate vs. calm	-0.094	0.006	246.735	1	0.000
	windy vs. moderate	-0.097	0.004	479.150	1	0.000
bm1	moderate vs. calm	-0.019	0.005	18.018	1	0.000
	windy vs. moderate	-0.032	0.004	63.340	1	0.000
bm2	moderate vs. calm	-0.032	0.005	46.148	1	0.000
	windy vs. moderate	-0.035	0.004	74.394	1	0.000
lskf	moderate vs. calm	-0.023	0.005	24.301	1	0.000
	windy vs. moderate	-0.036	0.004	81.429	1	0.000
lskm1	moderate vs. calm	-0.035	0.005	52.492	1	0.000
	windy vs. moderate	-0.041	0.004	94.754	1	0.000
lskm2	moderate vs. calm	-0.023	0.005	25.381	1	0.000
	windy vs. moderate	-0.030	0.004	51.035	1	0.000
mp	moderate vs. calm	-0.031	0.005	42.925	1	0.000
	windy vs. moderate	-0.049	0.004	149.117	1	0.000
trilH	moderate vs. calm	-0.015	0.005	10.758	1	0.001
	windy vs. moderate	-0.033	0.004	66.246	1	0.000
trilL	moderate vs. calm	0.002	0.004	0.131	1	0.717
	windy vs. moderate	-0.022	0.004	28.752	1	0.000
bittern	moderate vs. calm	-0.016	0.005	11.789	1	0.001
	windy vs. moderate	-0.001	0.004	0.048	1	0.827
kBoom	moderate vs. calm	-0.028	0.005	31.448	1	0.000
	windy vs. moderate	-0.011	0.005	5.909	1	0.015
kc	moderate vs. calm	0.000	0.004	0.001	1	0.978
	windy vs. moderate	-0.017	0.004	19.365	1	0.000
weka	moderate vs. calm	-0.045	0.005	81.488	1	0.000
	windy vs. moderate	-0.053	0.004	175.045	1	0.000
kaka	moderate vs. calm	-0.005	0.004	1.623	1	0.203
	windy vs. moderate	-0.030	0.003	80.155	1	0.000
hihi	moderate vs. calm	0.003	0.005	0.459	1	0.498
	windy vs. moderate	-0.032	0.004	73.159	1	0.000
robin	moderate vs. calm	-0.010	0.005	4.936	1	0.026
	windy vs. moderate	-0.025	0.004	41.229	1	0.000
tui	moderate vs. calm	-0.011	0.004	6.791	1	0.009
	windy vs. moderate	-0.032	0.004	67.643	1	0.000
sad1	moderate vs. calm	-0.010	0.005	4.861	1	0.027
	windy vs. moderate	-0.032	0.004	66.643	1	0.000
sad2	moderate vs. calm	0.000	0.005	0.001	1	0.981
	windy vs. moderate	-0.026	0.004	34.663	1	0.000
sad3	moderate vs. calm	-0.019	0.005	16.838	1	0.000
	windy vs. moderate	-0.029	0.004	54.368	1	0.000

Table A.14: EMM (Analysis III) – Wind level – overall test results – significant effects after sequential Sidak correction on $\alpha = 0.01$

Call example	Wald Chi-Square	df	Significance
bf	661.628	2	0.000
bm1	93.953	2	0.000
bm2	134.927	2	0.000
lskf	122.889	2	0.000
lskm1	162.494	2	0.000
lskm2	88.126	2	0.000
mp	221.131	2	0.000
trilH	87.483	2	0.000
trilL	29.810	2	0.000
bittern	14.525	2	0.001
kBoom	52.241	2	0.000
kc	20.499	2	0.000
weka	281.379	2	0.000
kaka	91.768	2	0.000
hihi	74.655	2	0.000
robin	51.044	2	0.000
tui	89.371	2	0.000
sad1	80.825	2	0.000
sad2	36.316	2	0.000
sad3	93.556	2	0.000

Bibliography

Hervé Abdi. The Bonferonni and Šidák corrections for multiple comparisons. *Encyclopedia of Measurement and Statistics*, 3:103–107, 2007. 3.2.7

Edward Aboufadel and Steven Schlicker. *Discovering Wavelets*. John Wiley & Sons, 2011. 4.4.3

Miguel A Acevedo and Luis J Villanueva-Rivera. Using automated digital recording systems as effective tools for the monitoring of birds and amphibians. *Wildlife Society Bulletin*, 34(1): 211–214, 2006. doi: http://dx.doi.org/10.2193/0091-7648(2006)34[211:UADRSA]2.0.CO;2. 2.4

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. doi: http://dx.doi.org/10.1016/j.ecoinf.2009.06.005. 2.3

Ian Agranat. Automatically identifying animal species from their vocalizations. In *Proceedings of the 5th International Conference on Bio-Acoustics, Holywell Park*, 2009. 2.3, 3.4.1

T Mitchell Aide, Carlos Corrada-Bravo, Marconi Campos-Cerqueira, Carlos Milan, Giovany Vega, and Rafael Alvarez. Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1:e103, 2013. doi: http://dx.doi.org/10.7717/peerj.103. 2.2.3, 2.3, 5.2.1

Kieran V Aland and Conrad J Hoskin. The advertisement call and clutch size of the Golden-capped Boulder-frog *Cophixalus pakayakulangun* (Anura: Microhylidae). *ZooTaxa*, 3718(3): 299–300, 2013. doi: http://dx.doi.org/10.11646/zootaxa.3718.3.8. 2.3

Pau Aleixandre, Julio Hernández Montoya, and Borja Milá. Speciation on Oceanic Islands: Rapid adaptive divergence vs. cryptic speciation in a Guadalupe Island songbird (Aves: *Junco*). *PloS one*, 8(5):e63242, 2013. doi: http://dx.doi.org/10.1371/journal.pone.0063242. 2.3

Renata D Alquezar and Ricardo B Machado. Comparisons between autonomous acoustic recordings and avian point counts in open woodland savanna. *The Wilson Journal of Ornithology*, 127(4):712–723, 2015. doi: http://dx.doi.org/10.1676/14-104.1. 2.4

Duncan Anderson, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, and Neeza Thandi. A practitioner's guide to Generalized Linear Models. *Casualty Actuarial Society Discussion Paper Program*, pages 1–116, 2004. 3.2.7

Sven E Anderson, Amish S Dave, and Daniel Margoliash. Template-based automatic recognition of birdsong syllables from continuous recordings. *The Journal of the Acoustical Society of America*, 100(2):1209–1219, 1996. doi: http://dx.doi.org/10.1121/1.415968. 2.2.5, 2.2.6, 5.1

Tórur Andreassen, Annemarie Surlykke, and John Hallam. Semi-automatic long-term acoustic surveying: A case study with bats. *Ecological Informatics*, 21:13–24, 2014. doi: http://dx.doi.org/10.1016/j.ecoinf.2013.12.010. 2.2.5, 2.3

George R Angehr, James Siegel, Constantino Aucca, Daniel G Christian, and Tatiana Pequeño. An assessment and monitoring program for birds in the Lower Urubamba Region, Peru. *Environmental Monitoring and Assessment*, 76(1):69–87, 2002. doi: http://dx.doi.org/10.1023/A:1015220921192. 2.1

J Edgardo Arévalo and Marcelo Araya-Salas. Collared forest-falcon (*Micrastur semitorquatus*) preying on chestnut-mandibled toucan (*Ramphastos swainsonii*) in Costa Rica. *The Wilson Journal of Ornithology*, 125(1):212–216, 2013. doi: http://dx.doi.org/10.1676/12-085.1. 2.3

Donald Aylor. Noise reduction by vegetation and ground. *The Journal of the Acoustical Society of America*, 51(1B):197–205, 1972. doi: http://dx.doi.org/10.1121/1.1912830. 3.1, 3.4.2

Myron C Baker and David M Logue. Population differentiation in a complex bird sound: A comparison of three bioacoustical analysis procedures. *Ethology*, 109(3):223–242, 2003. doi: http://dx.doi.org/10.1046/j.1439-0310.2003.00866.x. 2.2.3, 5.2.1

Myron C Baker and David M Logue. A comparison of three noise reduction procedures applied to bird vocal signals. *Journal of Field Ornithology*, 78(3):240–253, 2007. doi: http://dx.doi.org/10.1111/j.1557-9263.2007.00109.x. 2.1

Sarah Baldo and Daniel J Mennill. Vocal behavior of Great Curassows, a vulnerable neotropical bird. *Journal of Field Ornithology*, 82(3):249–258, 2011. doi: http://dx.doi.org/10.1111/j.1557-9263.2011.00328.x. 4

Rolf Bardeli, D Wolff, Frank Kurth, M Koch, K-H Tauchert, and K-H Frommolt. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, 31(12):1524–1534, 2010. doi: http://dx.doi.org/10.1016/j.patrec.2009.09.014. 2.2.5, 2.2, 2.4

Rosemary K Barraclough. Distance sampling: A discussion document produced for the Department of Conservation. *Department of Conservation Science & Research International Report 175*, 2000. URL `http://www.doc.govt.nz/documents/science-and-technical/ir175.pdf`. 1.1

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 *Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 1676–1679, May 2012. doi: http://dx.doi.org/10.1109/ISCAS.2012.6271580. 2.2.6, 2.3

SS Bee, J Pramod, and S Jilani. Real time speech denoising using simulink and beagle bone black. URL `http://www.ijcrce.org/paper_final/REAL%20TIME%20SPEECH%20DENOISING%20USING%20SIMULINK%20AND%20BEAGLE%20BONE%20BLACK.pdf`. Access date: 15/09/2015. 4.4.2

BirdLife International. Species factsheet: *Botaurus poiciloptilus*, 2016. URL `http://www.birdlife.org`. Access date: 29/08/2016. 5.2.1

Abraham L Borker, Portia Halbert, Matthew W Mckown, Bernie R Tershy, and Donald A Croll. A comparison of automated and traditional monitoring techniques for marbled murrelets using passive acoustic sensors. *Wildlife Society Bulletin*, 39(4):813–818, 2015. doi: http://dx.doi.org/10.1002/wsb.608. 2.4

Neil Boucher, Michihiro Jinnai, and Andrew Smolders. A fully automatic wildlife acoustic monitor and survey system. In *Proceedings of the Acoustics 2012 Nantes Conference*, April 2012. Access date: 02/05/2014. 2.2

Neil J Boucher. *SoundID Version 2.0.0 Documentation*. SoundID, 2014. URL `http://www.soundid.net/SoundID/Downloads/Running%20SoundID%202014%20Jan.pdf`. Access date: 24/01/2015. 2.2.1, 2.3

T Scott Brandes. Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International*, 18(S1):S163–S173, 2008. doi: http://dx.doi.org/10.1017/S0959270908000415. 2.1, 2.2.1, 2.3, 5.1

Forrest Briggs, Raviv Raich, and Xiaoli Z Fern. Audio classification of bird species: A statistical manifold approach. In *Proceedings of the 9th IEEE International Conference on Data Mining*, pages 51–60, Dec. 2009. doi: http://dx.doi.org/10.1109/ICDM.2009.65. 2.2.6

Forrest Briggs, Balaji Lakshminarayanan, Lawrence Neal, Xiaoli Z Fern, Raviv Raich, Sarah JK Hadley, Adam S Hadley, and Matthew G Betts. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131(6):4640–4650, 2012. doi: http://dx.doi.org/10.1121/1.4707424. 2.1, 2.2.5, 2.3, 5.1

Forrest Briggs, Yonghong Huang, Raviv Raich, Konstantinos Eftaxias, Zhong Lei, William Cukierski, Sarah Frey Hadley, Adam Hadley, Matthew Betts, Xiaoli Z Fern, et al. The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–8, Sept. 2013. doi: http://dx.doi.org/10.1109/MLSP.2013.6661934. 2.4

Alex Brighten. Vocalisations of the New Zealand morepork (*Ninox novaeseelandiae*) on Ponui Island. Master's thesis, Massey University, Palmerston North, New Zealand, 2015. 1.2, 2.1, 2.4, 5.2.1

Timothy J Brown and Paul Handford. Sound design for vocalizations: Quality in the woods, consistency in the fields. *The Condor*, 102(1):81–92, 2000. doi: http://dx.doi.org/10.1650/0010-5422(2000)102[0081:SDFVQI]2.0.CO;2. 3.4.1

Veronica L Bura, Vanya G Rohwer, Paul R Martin, and Jayne E Yack. Whistling in caterpillars (*Amorpha juglandis*, Bombycoidea): sound-producing mechanism and function. *Journal of Experimental Biology*, 214(1):30–37, 2011. doi: http://dx.doi.org/10.1242/jeb.046805. 2.3

C Sidney Burrus, Ramesh A Gopinath, and Haitao Guo. Generalizations of the basic multiresolution wavelet system. In *Introduction to Wavelets and Wavelet Transforms: A Primer*, pages 98–145. Prentice-Hall, Inc., 1997. 4.4.1

Rachel T Buxton and Ian L Jones. Measuring nocturnal seabird activity and status using acoustic recording devices: applications for island restoration. *Journal of Field Ornithology*, 83(1):47–60, 2012. doi: http://dx.doi.org/10.1111/j.1557-9263.2011.00355.x. 4.7

Rachel T Buxton, Heather L Major, Ian L Jones, and Jeffrey C Williams. Examining patterns in nocturnal seabird activity and recovery across the Western Aleutian Islands, Alaska, using automated acoustic recording. *The Auk*, 130(2):331–341, 2013. doi: http://dx.doi.org/10.1525/auk.2013.12134. 2.3

Samantha L Bye, Robert J Robel, and Kenneth E Kemp. Effects of human presence on vocalizations of grassland birds in Kansas. *The Prairie Naturalist*, 33(4):249–256, 2001. 2.1

Jinhai Cai, Dominic Ee, Binh Pham, Paul Roe, and Jinglan Zhang. Sensor network for the monitoring of ecosystem: Bird species recognition. In *Proceedings of the 3rd IEEE International Conference on Intelligent Sensors, Sensor Networks and Information*, pages 293–298, Dec. 2007. doi: http://dx.doi.org/10.1109/ISSNIP.2007.4496859. 2.2.7, 2.3

M. King Carolyn, Gaukrodger D. John, and A. Ritchie Neville. *The Drama of Conservation.* Springer, 2015. 1.1

Clive K Catchpole. Variation in the song of the great reed warbler *Acrocephalus arundinaceus* in relation to mate attraction and territorial defence. *Animal Behaviour*, 31(4):1217–1225, 1983. doi: http://dx.doi.org/10.1016/S0003-3472(83)80028-1. 4.2

Clive K Catchpole and Peter JB Slater. *Bird Song: Biological Themes and Variations.* Cambridge University Press, 2008. 4.2

RA Charif, LM Strickman, and AM Waack. *Raven Pro 1.4 User's Manual. Revision 11.* The Cornell Lab of Ornithology, Ithaca, NY, Dec. 2010. URL `http://www.birds.cornell.edu/brp/raven/Raven14UsersManual.pdf`. Access date: 29/08/2015. 2.3

Jingdong Chen, Jacob Benesty, Yiteng Huang, and Simon Doclo. New insights into the noise reduction Wiener filter. *IEEE Transactions on audio, speech, and language processing*, 14 (4):1218–1234, 2006. doi: http://dx.doi.org/10.1109/TSA.2005.860851. 2.2.3

Zhixin Chen and Robert C Maher. Semi-automatic classification of bird vocalizations using spectral peak tracks. *The Journal of the Acoustical Society of America*, 120(5):2974–2984, 2006. doi: http://dx.doi.org/10.1121/1.2345831. 2.2.5, 2.3, 2.2.7

Jinkui Cheng, Yuehua Sun, and Liqiang Ji. A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. *Pattern Recognition*, 43(11):3846–3852, 2010. doi: http://dx.doi.org/10.1016/j.patcog.2010.04.026. 4

Mei-Fang Cheng. The role of vocal self-stimulation in female responses to males: Implications for state-reading. *Hormones and Behavior*, 53(1):1–10, 2008. doi: http://dx.doi.org/10.1016/j.yhbeh.2007.08.007. 2.2

Chih-Hsun Chou and Pang-Hsin Liu. Bird species recognition by wavelet transformation of a section of birdsong. In *Proceedings of the IEEE Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing. UIC-ATC'09*, pages 189–193, July 2009. doi: http://dx.doi.org/10.1016/10.1109/UIC-ATC.2009.85. 2.2.6, 4.4.2

Wei Chu and Abeer Alwan. A correlation-maximization denoising filter used as an enhancement frontend for noise robust bird call classification. In *INTERSPEECH-2009*, pages 2831–2834, 2009. 2.1

Wei Chu and Abeer Alwan. FBEM: A filter bank EM algorithm for the joint optimization of features and acoustic model parameters in bird call classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1993–1996, March 2012. doi: http://dx.doi.org/10.1109/ICASSP.2012.6288298. 2.2.6

Wei Chu and Daniel T Blumstein. Noise robust bird song detection using syllable pattern-based hidden Markov models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 345–348, May 2011. doi: http://dx.doi.org/10.1109/ICASSP.2011.5946411. 2.3, 2.4, 5.1, 5.2.1, 5.3.4, 5.4

Christopher W Clark and Kurt M Fristrup. Advanced technologies for acoustic monitoring of bird populations. Technical report, Cornell University, Ithaca, NY, April 2009. 2.3

Patrick J Clemins and Michael T Johnson. Application of speech recognition to African elephant (*Loxodonta africana*) vocalizations. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–484–I–487, April 2003. doi: http://dx.doi.org/10.1109/ICASSP.2003.1198823. 2.2.6

Rogan Colbourne and Andrew Digby. Call rate behaviour of brown kiwi (*Apteryx mantelli*) and great spotted kiwi (*A. haastii*) in relation to temporal and environmental parameters. *Department of Conservation Research and Development Series 348*, 2016. 1.2, 2.1, 5.1

John Coleman. *Introducing speech and language processing*. Cambridge University Press, 2005. 2.2.6

Clément Cornec, Yves Hingrat, and Fanny Rybak. Individual signature in a lekking species: Visual and acoustic courtship parameters may help discriminating conspecifics in the houbara bustard. *ethology*, 120(7):726–737, 2014. doi: http://dx.doi.org/10.1111/eth.12244. 4.7

Kathryn A Cortopassi and Jack W Bradbury. The comparison of harmonically rich sounds using spectrographic cross-correlation and principal coordinates analysis. *Bioacoustics*, 11 (2):89–127, 2000. 5.3.3

Jenna L Cragg, Alan E Burger, and John F Piatt. Testing the effectiveness of automated acoustic sensors for monitoring vocal activity of marbled murrelets *Brachyramphus marmoratus*. *Marine Ornithology*, 43:151–160, 2015. 2.3, 3.5

Laura Crothers, Eben Gering, and Molly Cummings. Aposematic signal variation predicts male–male interactions in a polymorphic poison frog. *Evolution*, 65(2):599–605, 2011. doi: http://dx.doi.org/10.1111/j.1558-5646.2010.01154.x. 2.3

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. doi: http://dx.doi.org/10.1071/wr02062. 2.1

Arij Daou, Frank Johnson, Wei Wu, and Richard Bertram. A computational tool for automated large-scale analysis and measurement of bird-song syntax. *Journal of Neuroscience Methods*, 210(2):147–160, 2012. doi: http://dx.doi.org/10.1016/j.jneumeth.2012.07.020. 2.3

Ingrid Daubechies. *Ten Lectures on Wavelets*, volume 61. SIAM, 1992. 5.2.2

Deanna K Dawson and Murray G Efford. Bird population density estimated from acoustic signals. *Journal of Applied Ecology*, 46(6):1201–1209, 2009. doi: http://dx.doi.org/10.1111/j.1365-2664.2009.01731.x. 2.1, 5.1

DG Dawson and PC Bull. Counting birds in New Zealand forests. *Notornis*, 22(2):101–109, 1975. URL `http://notornis.osnz.org.nz/system/files/Notornis_22_2.pdf`. 2.1

Allan G de Oliveira, Thiago M Ventura, Todor D Ganchev, Josiel M de Figueiredo, Olaf Jahn, Marinez I Marques, and Karl-L Schuchmann. Bird acoustic activity detection based on morphological filtering of the spectrogram. *Applied Acoustics*, 98:34–42, 2015. doi: http://dx.doi.org/10.1016/j.apacoust.2015.04.014. 2.2, 2.3, 5.1

Jennifer M Dent and Laura E Molles. Call-based identification as a potential tool for monitoring Great Spotted Kiwi. *Emu*, 116(4):315–322, 2016. doi: http://dx.doi.org/10.1071/MU15079. 4

Department of Conservation. New Zealand birds: Native animal conservation. URL `http://www.doc.govt.nz/nature/native-animals/birds/`. Access date: 29/08/2016. 1.1, 2.1

Andrew Digby. *Whistling in the dark: an acoustic study of little spotted kiwi*. PhD thesis, Victoria University of Wellington, 2013. 3.4.2

Andrew Digby, Michael Towsey, Ben D Bell, and Paul D Teal. A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution*, 4(7): 675–683, 2013. doi: http://dx.doi.org/10.1111/2041-210X.12060. 2.1, 2.2.7, 2.3, 2.4, 3.5, 4.1

Xueyan Dong, Michael Towsey, Anthony Truskinger, Mark Cottman-Fields, Jinglan Zhang, and Paul Roe. Similarity-based birdcall retrieval from environmental audio. *Ecological Informatics*, 29:66–76, 2015. doi: http://dx.doi.org/10.1016/j.ecoinf.2015.07.007. 2.3

Allison J Doupe and Patricia K Kuhl. Birdsong and human speech: common themes and mechanisms. *Annual review of neuroscience*, 22(1):567–631, 1999. doi: http://dx.doi.org/10.1146/annurev.neuro.22.1.567. 2.2.7

Shufei Duan, Michael Towsey, Jinglan Zhang, Anthony Truskinger, Jason Wimmer, and Paul Roe. Acoustic component detection for automatic species recognition in environmental monitoring. In *Proceedings of the 7th international conference on intelligent sensors, sensor networks and information processing (ISSNIP)*, pages 514–519, Dec. 2011. doi: http://dx.doi.org/10.1109/ISSNIP.2011.6146597. 2.3, 4.2

Shufei Duan, Jinglan Zhang, Paul Roe, Jason Wimmer, Xueyan Dong, Anthony Truskinger, and Michael Towsey. Timed probabilistic automaton: a bridge between Raven and Song Scope for automatic species recognition. In *Proceedings of the 25th Innovative Applications of Artificial Intelligence Conference*, pages 1519–1524. AAAI, 2013. URL `http://www.aaai.org/ocs/index.php/IAAI/IAAI13/paper/view/6092`. 2.3, 4.2

Olivier Dufour, Thierry Artieres, Hervé Glotin, and Pascale Giraudet. Clusterized mel filter cepstral coefficients and support vector machines for bird song identification. In *Proceedings of the 1st workshop on Machine Learning for Bioacoustics*, volume 951, pages 89–93, 2013. 2.3

Coen PH Elemans. The singer and the song: The neuromechanics of avian sound production. *Current Opinion in Neurobiology*, 28:172–178, 2014. doi: http://dx.doi.org/10.1016/j.conb.2014.07.022. 2.2

G.P. Elliott. Kakapo : In Miskelly, C.M. (ed.) New Zealand Birds Online, 2013. URL `www.nzbirdsonline.org.nz`. Access date: 04/07/2016. 5.1, 5.2.1

John T Emlen and Michael J DeJong. Counting birds: The problem of variable hearing abilities (contando aves: El problema de la variabilidad en la capacidad auditiva). *Journal of Field Ornithology*, 63(1):26–31, 1992. 2.1, 5.1

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010. doi: http://dx.doi.org/10.1145/1873951.1874246. 2.3

Seppo Fagerlund. *Automatic recognition of bird species by their sounds*. PhD thesis, Helsinki University of technology, 2004. 2.2.5

Seppo Fagerlund. Bird species recognition using support vector machines. *EURASIP Journal on Applied Signal Processing*, 2007(1):64–64, 2007. doi: http://dx.doi.org/10.1155/2007/38637. 2.3

Almo Farina. *Soundscape ecology: principles, patterns, methods and applications*. Springer, 2014. 3.1, 4.3.1

Chique M Fernandez. The real song of the sirens - a study assessing the variability in manatee vocalizations. Master's thesis, University of St Andrews, 2012. 2.3

Luiza Figueira, José L Tella, Ulisses M Camargo, and Gonçalo Ferraz. Autonomous sound monitoring shows higher use of Amazon old growth than secondary forest by parrots. *Biological Conservation*, 184:27–35, 2015. doi: http://dx.doi.org/10.1016/j.biocon.2014.12.020. 2.4

Gábor Fodor. The ninth annual MLSP competition: first place. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–2, Sept. 2013. doi: http://dx.doi.org/10.1109/MLSP.2013.6661932. 2.2.5, 2.2.7, 2.3, 5.1

Elizabeth JS Fox, J Dale Roberts, and Mohammed Bennamoun. Text-independent speaker identification in birds. In *Proceedings of the Interspeech 2006 – 9th International Conference on Spoken Language Processing*, 2006. 2.2.3, 2.2.6, 2.3, 5.2.1

Elizabeth JS Fox, J Dale Roberts, and Mohammed Bennamoun. Call-independent individual identification in birds. *Bioacoustics*, 18(1):51–67, 2008. doi: http://dx.doi.org/10.1080/09524622.2008.9753590. 2.1, 2.2.5, 5.1

Andreas Franzen and Irene YH Gu. Classification of bird species by using key song searching: a comparative study. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2003.*, volume 1, pages 880–887, Oct. 2003. doi: http://dx.doi.org/10.1109/ICSMC.2003.1243926. 2.2.5, 5.1

Karl-Heinz Frommolt and Klaus-Henry Tauchert. Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird. *Ecological Informatics*, 21:4–12, 2014. doi: http://dx.doi.org/10.1016/j.ecoinf.2013.12.009. 2.2.5, 2.2, 2.2.7, 2.3, 2.4

Brett J Furnas and Richard L Callas. Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. *The Journal of Wildlife Management*, 79(2):325–337, 2015. doi: http://dx.doi.org/10.1002/jwmg.821. 2.4

Todor Ganchev, Iosif Mporas, Olaf Jahn, Klaus Riede, Karl-L Schuchmann, and Nikos Fakotakis. Acoustic bird activity detection on real-field data. In *Hellenic Conference on Artificial Intelligence*, pages 190–197. Springer, 2012. doi: http://dx.doi.org/10.1007/978-3-642-30448-4_24. 2.3

Todor D Ganchev, Olaf Jahn, Marinez Isaac Marques, Josiel Maimone de Figueiredo, and Karl-L Schuchmann. Automated acoustic detection of Vanellus chilensis lampronotus. *Expert Systems with Applications*, 42(15):6098–6111, 2015. doi: http://dx.doi.org/10.1016/j.eswa.2015.03.036. 2.3, 2.4

Gillian Gilbert, Peter K McGregor, and Glen Tyler. Vocal individuality as a census tool: Practical considerations illustrated by a study of two rare species (individualidad vocal como herramienta en los censos: Consideraciones prácticas ilustradas por un estudio de dos especies raras. *Journal of Field Ornithology*, 65(3):335–348, 1994. URL `http://www.jstor.org/stable/4513949`. 4

H Glotin, Y LeCun, T Artiéres, S Mallat, O Tchernichovski, and X Halkias. Neural information processing scaled for bioacoustics, from neurons to big data. In *Proceedings of NIPS4B, international workshop joint to NIPS*, 2013. URL `http://sabiod.org/NIPS4B2013_book.pdf`. Access date: 13/04/2016. 2.4

Michelle Goh. Developing an automated acoustic monitoring system to estimate abundance of Corys Shearwaters in the Azores. Master's thesis, Imperial College London, 2011. 2.3

Jennifer L Goyette, Robert W Howe, Amy T Wolf, and W Douglas Robinson. Detecting tropical nocturnal birds using automated audio recordings. *Journal of Field Ornithology*, 82(3):279–287, 2011. doi: http://dx.doi.org/10.1111/j.1557-9263.2011.00331.x. 2.1

Martin Graciarena, Michelle Delplanche, Elizabeth Shriberg, Andreas Stolcke, and Luciana Ferrer. Acoustic front-end optimization for bird species recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 293–296, March 2010. doi: http://dx.doi.org/10.1109/ICASSP.2010.5495923. 2.2.6, 2.2.6, 2.3

Amara Graps. An introduction to wavelets. *IEEE computational science and engineering*, 2 (2):50–61, 1995. doi: http://dx.doi.org/10.1109/99.388960. 4.4, 5.2.2

Richard D Gregory, David W Gibbons, and Paul F Donald. Bird census and survey techniques. In *Bird Ecology and Conservation*, pages 17–56. Oxford University Press, 2004. 5.1

Donald R Griffin. The importance of atmospheric attenuation for the echolocation of bats (Chiroptera). *Animal Behaviour*, 19(1):55–61, 1971. doi: http://dx.doi.org/10.1016/S0003-3472(71)80134-3. 3.4.1

Berke M Gur and Christopher Niezrecki. Autocorrelation based denoising of manatee vocalizations using the undecimated discrete wavelet transform. *The Journal of the Acoustical Society of America*, 122(1):188–199, 2007. doi: http://dx.doi.org/10.1121/1.2735111. 4.4.2

Jonathan T Hagstrum. Infrasound and the avian navigational map. *Journal of Experimental Biology*, 203(7):1103–1111, 2000. URL `http://jeb.biologists.org/content/jexbio/203/7/1103.full.pdf`. Access date: 12/04/2015. 4.7

A Harma and Panu Somervuo. Classification of the harmonic structure in bird vocalization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. (ICASSP'04)*, volume 5, pages V–701–4, May 2004. doi: http://dx.doi.org/10.1109/ICASSP.2004.1327207. 2.2.5, 2.2, 5.1

Aki Harma. Automatic identification of bird species based on sinusoidal modeling of syllables. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. (ICASSP'03)*, volume 5, pages V–545–8, April 2003. doi: http://dx.doi.org/10.1109/ICASSP.2003.1200027. 2.2, 2.3

G.A. Harper. Kakapo on Little Barrier (Hauturu) Island. Annual report for the year July 1994-June 1995. Science & Research Internal Report 160, 1998. URL `http://www.doc.govt.nz/Documents/science-and-technical/SRIR160.pdf`. Access date: 29/08/2016. 5.2.1

John Haselmayer and James S Quinn. A comparison of point counts and sound recording as bird survey methods in Amazonian southeast Peru. *The Condor*, 102(4):887–893, 2000. doi: http://dx.doi.org/10.1650/0010-5422(2000)102[0887:ACOPCA]2.0.CO;2. 2.4, 5.1

Jason R Heller and John D Pinezich. Automatic recognition of harmonic bird sounds using a frequency track extraction algorithm. *The Journal of the Acoustical Society of America*, 124 (3):1830–1837, 2008. doi: http://dx.doi.org/10.1121/1.2950085. 2.3

Samuel D Hill, Weihong Ji, Kevin A Parker, Christophe Amiot, and Sarah J Wells. A comparison of vocalisations between mainland tui (*Prosthemadera novaeseelandiae novaeseelandiae*) and chatham island tui (*P. n. chathamensis*). *New Zealand Journal of Ecology*, 37(2):214–223, 2013. URL `http://www.jstor.org/stable/24060784`. 4

Keith A Hobson, Robert S Rempel, Hamilton Greenwood, Brian Turnbull, and Steven L Van Wilgenburg. Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. *Wildlife Society Bulletin*, 30(3):709–720, 2002. URL `http://www.jstor.org/stable/3784223`. 2.4

Stephen B Holmes, Kenneth A McIlwrick, and Lisa A Venier. Using automated sound recording and analysis to detect bird species-at-risk in southwestern Ontario woodlands. *Wildlife Society Bulletin*, 38(3):591–598, 2014. doi: http://dx.doi.org/10.1002/wsb.421. 2.4

Alain Hore and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *Proceedings of the 20th International Conference on Pattern Recognition*, pages 2366–2369. IEEE, Aug. 2010. doi: http://dx.doi.org/10.1109/ICPR.2010.579. 4.5.2

Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008. doi: http://dx.doi.org/10.1049/el:20080522. 4.5.2

Uno Ingård. A review of the influence of meteorological conditions on sound propagation. *The Journal of the Acoustical Society of America*, 25(3):405–411, 1953. doi: http://dx.doi.org/10.1121/1.1907055. 3.1, 3.4.1, 3.4.2

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):1–10, 2011. doi: http://dx.doi.org/10.1155/2011/982936. 2.3, 2.2.7

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. doi: http://dx.doi.org/10.1109/LSP.2015.2409173. 2.2.5, 2.2, 2.3, 2.2.7

Michihiro Jinnai, Neil Boucher, Minoru Fukumi, and Hollis Taylor. A new optimization method of the geometric distance in an automatic recognition system for bird vocalisations. In *Acoustics 2012*, 2012. 2.2.5, 2.2, 2.2.7, 2.3, 5.1, 5.2.2, 5.2.3, 5.3.3

Chia-Feng Juang and Tai-Mou Chen. Birdsong recognition using prediction-based recurrent neural fuzzy networks. *Neurocomputing*, 71(1):121–130, 2007. doi: http://dx.doi.org/10.1016/j.neucom.2007.08.011. 2.2.5, 2.2, 2.3, 5.1

Eric P Kasten, Philip K McKinley, and Stuart H Gage. Ensemble extraction for classification and detection of bird species. *Ecological Informatics*, 5(3):153–166, 2010. doi: http://dx.doi.org/10.1016/j.ecoinf.2010.02.003. 2.3

Jonathan Katz, Sasha D Hafner, and Therese Donovan. Assessment of error rates in acoustic monitoring with the R package monitoR. *Bioacoustics*, 25(2):1–20, 2016. doi: http://dx.doi.org/10.1080/09524622.2015.1133320. 2.2.1

Marten Ken, Quine Douglas, and Marler Peter. Sound transmission and its significance for animal vocalization: II. Tropical forest habitats. *Behavioral Ecology and Sociobiology*, 2(3):291–302, 1977. ISSN 03405443, 14320762. doi: http://dx.doi.org/10.1007/BF00299741. 3.1, 3.4.1, 3.4.2

Lawrence E Kinsler and Austin Rogers Frey. *Fundamentals of acoustics*. New York, Wiley, 2nd edition, 1962. 3.1

Alexander NG Kirschel, Martin L Cody, Zachary T Harlow, Vasilis J Promponas, Edgar E Vallejo, and Charles E Taylor. Territorial dynamics of Mexican Ant-thrushes Formicarius moniliger revealed by individual recognition of their songs. *IBIS*, 153(2):255–268, 2011. doi: http://dx.doi.org/10.1111/j.1474-919X.2011.01102.x. 2.3

Brian T Klingbeil and Michael R Willig. Bird biodiversity assessments in temperate forest: the value of point count versus acoustic monitoring protocols. *PeerJ*, 3:e973, 2015. doi: https://doi.org/10.7717/peerj.973. 2.4

Joseph A Kogan and Daniel Margoliash. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study. *The Journal of the Acoustical Society of America*, 103(4):2185–2196, 1998. doi: http://dx.doi.org/10.1121/1.421364. 2.2.6, 2.3, 2.2.7

Donald E Kroodsma. *The Singing Life of Birds: The Art and Science of Listening to Birdsong.* Boston: Houghton Mifflin Harcourt, 2005. 4.2

Donald E Kroodsma, Edward H Miller, and Henri Ouellet. *Acoustic Communication in Birds*. New York: Academic Press, 1982. URL `https://books.google.co.nz/books?id=6HgEDQEACAAJ`. 4.2

Chiman Kwan, Gang Mei, X Zhao, Zhubing Ren, Roger Xu, Vincent Stanford, Cedric Rochet, Julian Aube, and KC Ho. Bird classification algorithms: Theory and experimental results. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004*, volume 5, pages V–289, May 2004. doi: http://dx.doi.org/10.1109/ICASSP.2004.1327104. 2.2.7

Chiman Kwan, KC Ho, Gang Mei, Yunhong Li, Zhubing Ren, Roger Xu, Y Zhang, Debang Lao, M Stevenson, Vincent Stanford, and Cedric Rochet. An automated acoustic system to monitor and classify birds. *EURASIP Journal on Advances in Signal Processing*, 2006(1): 1–19, 2006. doi: http://dx.doi.org/10.1155/ASP/2006/96706. 2.3

HJ Landau. Sampling, data transmission, and the Nyquist rate. *Proceedings of the IEEE*, 55 (10):1701–1706, 1967. doi: http://dx.doi.org/10.1109/PROC.1967.5962. 2.2.1

Mario Lasseck. Bird song classification in field recordings: winning solution for NIPS4B 2013 competition. In *Proceedings of the International symposium Neural Information Scaled for Bioacoustics, joint to NIPS, Nevada*, pages 176–181, 2013. URL `http://sabiod.org/nips4b`. 2.1, 2.2, 2.3, 5.1, 5.2.3, 5.3.3

Mario Lasseck. Large-scale identification of birds in audio recordings. In *Working Notes of CLEF 2014*, pages 643–653, 2014. 2.2.5, 2.3

Mario Lasseck. Improved automatic bird identification through decision tree based feature selection and bagging. In *Working notes of CLEF 2015 conference*, 2015a. 2.2.7, 2.3, 2.2.7, 5.1

Mario Lasseck. Towards automatic large-scale identification of birds in audio recordings. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, volume 9283, pages 364–375. Springer, 2015b. doi: http://dx.doi.org/10.1007/978-3-319-24027-5_39. 5.1, 5.2.3, 5.3.3

Chang-Hsing Lee, Yeuan-Kuen Lee, and Ren-Zhuang Huang. Automatic recognition of bird songs using cepstral coefficients. *Journal of Information Technology and Applications*, 1(1): 17–23, 2006. 2.2, 2.2.6, 2.3

Chang-Hsing Lee, Chin-Chuan Han, and Ching-Chien Chuang. Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1541–1550, 2008. doi: http://dx.doi.org/10.1109/TASL.2008.2005345. 2.3

Chang-Hsing Lee, Sheng-Bin Hsu, Jau-Ling Shih, and Chih-Hsun Chou. Continuous birdsong recognition using gaussian mixture modeling of image shape features. *IEEE Transactions on Multimedia*, 15(2):454–464, 2013. doi: http://dx.doi.org/10.1109/TMM.2012.2229969. 2.3

M Ross Lein. Territorial and courtship songs of birds. *Nature*, 237:48 – 49, 1972. doi: http://dx.doi.org/10.1038/237048a0. 4.2

Christophe Lévy, Georges Linarès, and Pascal Nocera. Comparison of several acoustic modeling techniques and decoding algorithms for embedded speech recognition systems. In *Workshop on DSP in Mobile and Vehicular Systems, Nagoya, Japan*. CiteSeer, 2003. 2.2.6

HF Linskens, MJM Martens, HJGM Hendriksen, AM Roestenberg-Sinnige, WAJM Brouwers, ALHC Van Der Staak, and AMJ Strik-Jansen. The acoustic climate of plant communities. *Oecologia*, 23(3):165–177, 1976. doi: http://dx.doi.org/10.1007/BF00361233. 3.4.2

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 *Proceedings of the 2011 IEEE International Symposium on Multimedia*, pages 117–122, Dec. 2011a. doi: http://dx.doi.org/10.1109/ISM.2011.27. 2.2.7, 2.3

Marcelo Teider Lopes, Carlos Nascimento Silla Junior, Alessandro Lameiras Koerich, and Celso Antonio Alves Kaestner. Feature set comparison for automatic bird species identification. In *Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics*, pages 965–970, Oct. 2011b. doi: http://dx.doi.org/10.1109/ICSMC.2011.6083794. 2.3

Richard H Loyn. *The 20-minute search: a simple method for counting forest birds*. Biological Survey Branch, State Forests and Lands Service, 1985. 2.1

Hong-feng Ma, Jian-wu Dang, and Xin Liu. Research of the optimal wavelet selection on entropy function. In *Future Control and Automation*, pages 35–42. Springer, 2012. doi: http://dx.doi.org/10.1007/978-3-642-31003-4_5. 4.4.1

X Ma, Chengke Zhou, and IJ Kemp. Automated wavelet selection and thresholding for PD detection. *IEEE Electrical Insulation Magazine*, 18(2):37–45, 2002. doi: http://dx.doi.org/10.1109/57.995398. 4.4.2, 4.4.3

Peter Maciej, Julia Fischer, and Kurt Hammerschmidt. Transmission characteristics of primate vocalizations: implications for acoustic analyses. *PloS one*, 6(8):e23015, 2011. doi: http://dx.doi.org/10.1371/journal.pone.0023015. 3.4.1, 3.4.2

N Madhu. Note on measures for spectral flatness. *Electronics Letters*, 45(23):1195–1196, 2009. doi: http://dx.doi.org/10.1049/el.2009.1977. 2.3

John Makhoul and Richard Schwartz. State of the art in continuous speech recognition. *Proceedings of the National Academy of Sciences*, 92(22):9956–9963, oct. 1995. 2.2.6

Stephen Marsland. *Machine Learning: An Algorithmic Perspective.* Chapman and Hall/ CRC, 2nd edition, 2014. 2.2.4, 2.2.6, 2.2.7, 4.4.1

Ken Marten and Peter Marler. Sound transmission and its significance for animal vocalization. I. Temporate habitats. *Behavioral Ecology and Sociobiology*, 2(3):271–290, 1977. doi: http://dx.doi.org/10.1007/BF00299740. 3.1, 3.2.6, 3.4.1, 3.4.2

Eiji Matsunaga and Kazuo Okanoya. Evolution and diversity in avian vocal system: An Evo-Devo model from the morphological and behavioral perspectives. *Development, Growth & Differentiation*, 51(3):355–367, 2009. doi: http://dx.doi.org/10.1111/j.1440-169X.2009.01091.x. 2.2

Alex L McIlraith and Howard C Card. Birdsong recognition using backpropagation and multivariate statistics. *IEEE Transactions on Signal Processing*, 45(11):2740–2748, 1997. doi: http://dx.doi.org/10.1109/78.650100. 2.3

Margaret A McLaren and Michael D Cadman. Can novice volunteers provide credible data for bird surveys requiring song identification? *Journal of Field Ornithology*, 70(4):481–490, 1999. 2.1

Alfred Mertins. Wavelet Transform. In *Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications*, chapter 8, pages 210–264. John Wiley & Sons, 2001. doi: http://dx.doi.org/10.1002/0470841834.ch8. 4.4

GB Mindlin. The physics of birdsong production. *Contemporary Physics*, 54(2):91–96, 2013. doi: http://dx.doi.org/10.1080/00107514.2013.810852. 2.2.1

Colin M Miskelly, John E Dowding, Graeme P Elliott, Rodney A Hitchmough, Ralph G Powlesland, Hugh A Robertson, Paul M Sagar, R Paul Scofield, and Graeme A Taylor. Conservation status of New Zealand birds, 2008. *Notornis*, 55(3):117–135, 2008. 1.1, 2.4

David Moffat, David Ronan, and Joshua D Reiss. An evaluation of audio feature extraction toolboxes. In *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15), Trondheim, Norway*, Dec 2015. 2.2.6

Jean Morlet, G Arens, E Fourgeau, and D Glard. Wave propagation and sampling theory-part I: Complex signal and scattering in multilayered media. *Geophysics*, 47(2):203–221, 1982. doi: http://dx.doi.org/10.1190/1.1441328. 4.4

Douglass H Morse. Territorial and courtship songs of birds. *Nature*, 226:659–661, 1970. doi: http://dx.doi.org/10.1038/226659a0. 4.2

Eugene S Morton. Ecological sources of selection on avian sounds. *American Naturalist*, 109 (965):17–34, 1975. doi: http://dx.doi.org/10.1086/282971. 3.1, 3.4.1, 3.4.2

Lindasalwa Muda, Mumtaj Begam, and I Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *Journal of Computing*, 2(3), March 2010. 2.2.6

Roger Mundry and Christian Sommer. Tonal vocalizations in a noisy environment: an approach to their semi-automatic analysis and examples of its application. *Anais da Academia Brasileira de Ciências*, 76(2):284–288, 2004. doi: http://dx.doi.org/10.1590/s0001-37652004000200016. 2.3

Rafael Hernández Murcia and Víctor Suárez Paniagua. Bird identification from continuous audio recordings. In *Proceedings of the 1st Workshop on Machine Learning for Bioacoustics joint to the 30th International Conference on Machine Learning (ICML 2013)*, pages 96–97, June 2013. 2.2.5, 2.3

Abbas Najafipour, Abbas Babaee, and S Mohammad Shahrtash. Comparing the trustworthiness of signal-to-noise ratio and peak signal-to-noise ratio in processing noisy partial discharge signals. *IET Science, Measurement & Technology*, 7(2):112–118, 2013. doi: http://dx.doi.org/10.1049/iet-smt.2012.0113. 4.5.2

Lawrence Neal, Forrest Briggs, Raviv Raich, and Xiaoli Z Fern. Time-frequency segmentation of bird song in noisy acoustic environments. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2012–2015, May 2011. doi: http://dx.doi.org/10.1109/ICASSP.2011.5946906. 2.2

Colin FJ O'Donnell and Emma M Williams. Protocols for the inventory and monitoring of populations of the endangered Australasian bittern (*Botaurus poiciloptilus*) in New Zealand. *Department of Conservation Technical Series 38*, 2015. (document), 5.1, 5.2.1, 5.8, 6.3

Colin FJ O'Donnell, Emma M Williams, and John Cheyne. Close approaches and acoustic triangulation: techniques for mapping the distribution of booming Australasian bittern (*Botaurus poiciloptilus*) on small wetlands. *Notornis*, 60(4):279–284, 2013. 5.2.1

AL O'Loghlen and MD Beecher. Mate, neighbour and stranger songs: a female song sparrow perspective. *Animal Behaviour*, 58(1):13–20, 1999. doi: http://dx.doi.org/10.1006/anbe.1999.1125. 2.2

Timos Papadopoulos, Stephen Roberts, and Kathy Willis. Detecting bird sound in unknown acoustic background using crowdsourced training data. *ArXiv e-prints arXiv:1505.06443*, 2015. 2.3, 2.2.7

Suvarna Patil, Gajendra Singh Chandel, and Ravindra Gupta. Performance analysis of steganography based on 5-wavelet families by 4 levels-DWT. *International Journal of Computer Science and Network Security (IJCSNS)*, 14(12):56–61, 2014. 4.4

Michael Syskind Pedersen, Jan Larsen, Ulrik Kjems, and Lucas C Parra. A survey of convolutive blind source separation methods. *Multichannel Speech Processing Handbook*, pages 1065–1084, 2007. URL `http://www.iro.umontreal.ca/~pift6080/H08/documents/papers/blind_source_tutorial.pdf`. 4.3.1

David J Perkel and Michael A Farries. Complementary 'bottom-up' and 'top-down' approaches to basal ganglia function. *Current Opinion in Neurobiology*, 10(6):725–731, 2000. doi: http://dx.doi.org/10.1016/S0959-4388(00)00156-2. 2.2

Bryan C Pijanowski, Almo Farina, Stuart H Gage, Sarah L Dumyahn, and Bernie L Krause. What is soundscape ecology? An introduction and overview of an emerging new science. *Landscape Ecology*, 26(9):1213–1232, 2011a. doi: http://dx.doi.org/10.1007/s10980-011-9600-8. 3.1

Bryan C Pijanowski, Luis J Villanueva-Rivera, Sarah L Dumyahn, Almo Farina, Bernie L Krause, Brian M Napoletano, Stuart H Gage, and Nadia Pieretti. Soundscape ecology: the science of sound in the landscape. *BioScience*, 61(3):203–216, 2011b. doi: http://dx.doi.org/10.1525/bio.2011.61.3.6. 2.2.3, 3.1

Carina Pohnke, Alison Evans, and MH Bowie. Morepork (*Ninox novaseelandiae*) distribution and conservation on Banks Peninsula. Technical report, Lincoln University, May 2015. 5.2.1

Ilyas Potamitis. Automatic classification of a taxon-rich community recorded in the wild. *PloS one*, 9(5):e96936, 2014. doi: http://dx.doi.org/10.1371/journal.pone.0096936. 2.1, 2.2.3, 2.2.3, 2.2.5, 2.2.7, 2.3, 2.4, 5.1

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. doi: http://dx.doi.org/10.1016/j.apacoust.2014.01.001. 2.1, 2.2.3, 2.2, 2.3, 3.5, 4.1, 5.1, 5.3.4, 5.4

Nirosha Priyadarshani, Stephen Marsland, Isabel Castro, and Amal Punchihewa. Birdsong denoising using wavelets. *PloS one*, 11(1):e0146790, 2016. doi: http://dx.doi.org/10.1371/journal.pone.0146790. 2.2.3, 2.1, 2.2.6, 5.2.2, 5.2.2

PGN Priyadarshani, NGJ Dias, and Amal Punchihewa. Dynamic time warping based speech recognition for isolated sinhala words. In *Proceedings of the IEEE 55th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 892–895, Aug. 2012. doi: http://dx.doi.org/10.1109/MWSCAS.2012.6292164. 2.2.6, 4.7

Ladislav Ptacek, Lukas Machlica, Pavel Linhart, Pavel Jaska, and Ludek Muller. Automatic recognition of bird individuals on an open set using as-is recordings. *Bioacoustics*, 25(1): 55–73, 2016. doi: http://dx.doi.org/10.1080/09524622.2015.1089524. 4

Richard Ranft. Natural sound archives: past, present and future. *Anais da Academia Brasileira de Ciências*, 76(2):456–460, 2004. doi: http://dx.doi.org/10.1590/S0001-37652004000200041. 2.4

Louis Ranjard, Sarah J Withers, Dianne H Brunton, Howard A Ross, and Stuart Parsons. Integration over song classification replicates: Song variant analysis in the hihi. *The Journal of the Acoustical Society of America*, 137(5):2542–2551, 2015. doi: http://dx.doi.org/10.1121/1.4919329. 2.2

Dustin G Reichard and Rindy C Anderson. Why signal softly? the structure, function and evolutionary significance of low-amplitude signals. *Animal Behaviour*, 105:253–265, 2015. doi: http://dx.doi.org/10.1016/j.anbehav.2015.04.017. 2.2

Yao Ren, Michael T Johnson, and Jidong Tao. Perceptually motivated wavelet packet transform for bioacoustic signal enhancement. *The Journal of the Acoustical Society of America*, 124 (1):316–327, 2008. doi: http://dx.doi.org/10.1121/1.2932070. 2.1, 2.2.7, 4.4.2, 4.7

Douglas G Richards and R Haven Wiley. Reverberations and amplitude fluctuations in the propagation of sound in a forest: implications for animal communication. *The American Naturalist*, 115(3):381–399, 1980. 3.1, 3.4.1, 3.4.1, 3.4.3

H A Robertson. North island brown kiwi: *In* Miskelly, C.M. (ed.) New Zealand Birds Online, 2013. URL `http://nzbirdsonline.org.nz/species/north-island-brown-kiwi`. Access date: 15/10/2016. 5.1, 5.2.1

Steven S Rosenstock, David R Anderson, Kenneth M Giesen, Tony Leukering, Michael F Carter, and F Thompson III. Landbird counting techniques: current practices and an alternative. *The Auk*, 119(1):46–53, 2002. doi: http://dx.doi.org/10.1642/0004-8038(2002)119[0046:LCTCPA]2.0.CO;2. 2.1

Derek J Ross. Bird call recognition with artificial neural networks, support vector machines, and kernel density estimation. Master's thesis, Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada, 2006. 2.3

Stephen I Rothstein and Robert C Fleischer. Vocal dialects and their possible relation to honest status signalling in the brown-headed cowbird. *The Condor*, 89(1):1–23, 1987. doi: http://dx.doi.org/10.2307/1368756. 3.4.1

Mareile Große Ruse, Dennis Hasselquist, Bengt Hansson, Maja Tarka, and Maria Sandsten. Automated analysis of song structure in complex birdsongs. *Animal Behaviour*, 112:39–51, 2016. doi: http://dx.doi.org/10.1016/j.anbehav.2015.11.013. 2.2.1

Md Sahidullah and Goutam Saha. Comparison of speech activity detection techniques for speaker recognition. *CoRR*, abs/1210.0297, 2012. 2.3

Luis Sandoval and Gilbert Barrantes. Characteristics of male spot-bellied bobwhite (*Colinus leucopogon*) song during territory establishment. *Journal of Ornithology*, 153(2):547–554, 2012. doi: http://dx.doi.org/10.1007/s10336-011-0775-1. 2.3

John R Sauer, Bruce G Peterjohn, and William A Link. Observer differences in the north american breeding bird survey. *The Auk*, 111(1):50–62, 1994. doi: http://dx.doi.org/10.2307/4088504. 2.1, 5.1

A Saunders and DA Norton. Ecological restoration at mainland islands in New Zealand. *Biological Conservation*, 99(1):109–119, 2001. doi: http://dx.doi.org/10.1016/s0006-3207(00)00192-0. 2.4

R Murray Schafer. *The tuning of the world*. Alfred A Knopf, 1977. 3.1

Thijs Schrama, Martin Poot, Magnus Robb, and Hans Slabbekoorn. Automated monitoring of avian flight calls during nocturnal migration. In *Proceedings of the International Expert meeting on IT-based detection of bioacoustical patterns, Computational bioacoustics for assessing biodiversity*, pages 131–134, Dec 2007. 2.2.3, 2.2, 5.1

Esther Sebastián-González, Joshua Pang-Ching, Jomar M Barbosa, and Patrick Hart. Bioacoustics for species management: two case studies with a hawaiian forest bird. *Ecology and evolution*, 5(20):4696–4705, 2015. doi: http://dx.doi.org/10.1002/ece3.1743. 2.3

Ondřej Sedláček, Jana Vokurková, Michal Ferenc, Eric Nana Djomo, Tomáš Albrecht, and David Hořák. A comparison of point counts with a new acoustic sampling method: a case study of a bird community from the montane forests of mount cameroon. *Ostrich*, 86(3):213–220, 2015. doi: http://dx.doi.org/10.2989/00306525.2015.1049669. 2.4

Arja Selin, Jari Turunen, and Juha T Tanttu. Wavelets in recognition of bird sounds. *EURASIP Journal on Applied Signal Processing*, 2007(1):141–141, 2007. doi: http://dx.doi.org/10.1155/2007/51806. 2.1, 2.2.5, 2.2.6, 2.2.7, 2.3, 4.4.2, 4.7, 5.1

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423., 1948. doi: http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x. 4.4.1

Raghavendra Sharma and Vuppuluri Prem Pyara. A robust denoising algorithm for sounds of musical instruments using wavelet packet transform. *Circuits and Systems*, 4(7):459–465, 2013. doi: http://dx.doi.org/10.4236/cs.2013.47060. 4.4.2

Ivy Shim, John J Soraghan, and WH Siew. Detection of PD utilizing digital signal processing methods. part 3: Open-loop noise reduction. *IEEE Electrical Insulation Magazine*, 17(1):6–13, 2001. doi: http://dx.doi.org/10.1109/57.901611. 4.4.2, 4.4.3

Nilu Singh, RA Khan, and Raj Shree. MFCC and prosodic feature extraction techniques: A comparative study. *International Journal of Computer Applications*, 54(1):9–13, 2012. 2.2.6

Mark D Skowronski and John G Harris. Acoustic detection and classification of microchiroptera using machine learning: lessons learned from automatic speech recognition. *The Journal of the Acoustical Society of America*, 119(3):1817–1833, 2006. doi: http://dx.doi.org/10.1121/1.2166948. 2.2.7

Panu Somervuo and Aki Härmä. Analyzing bird song syllables on the self-organizing map. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, 2003. 2.2.6, 2.3

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. doi: http://dx.doi.org/10.1109/TASL.2006.872624. 2.2.5, 2.2, 2.2.6, 2.2.7, 2.3, 4.2, 5.1

R Specht. AVISOFT– Sound analysis and synthesis laboratory Pro: a PC program for sonagraphic analysis. *AVISOFT, Berlin, Germany*, 1993. 2.2.5, 2.3

Erick Stattner, Nicolas Vidot, Philippe Hunel, and Martine Collard. Wireless sensor network for habitat monitoring: A counting heuristic. In *Proceedings of the IEEE 37th Conference on Local Computer Networks Workshops (LCN Workshops)*, pages 753–760, Oct. 2012. doi: http://dx.doi.org/10.1109/MLSP.2013.6661934. 1.2, 2.1

Erick Stattner, Wilfried Segretier, Martine Collard, Philippe Hunel, and Nicolas Vidot. Song-based classification techniques for endangered bird conservation. In ICML 2013 Workshop on Machine Learning for Bioacoustics, Atlanta, Georgia, USA, 2013. URL `http://arxiv.org/abs/1306.5349`. 2.2.6

Rebecca L Stirnemann, Murray A Potter, David Butler, and Edward O Minot. Acoustic differences enable sex discrimination in Ma'oma'o (*Gymnomyza samoensis*), a species with high sexual morphological overlap. *The Wilson Journal of Ornithology*, 127(3):376–386, 2015. 2.4

Dan Stowell and Mark D Plumbley. Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers. Technical report, Centre for Digital Music, Queen Mary, University of London, July 2011. 2.3

Dan Stowell and Mark D Plumbley. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, 2:e488, 2014. doi: https://doi.org/10.7717/peerj.488. 2.2.7

William J Sutherland, Ian Newton, and Rhys Green. *Bird Ecology and Conservation: A Handbook of Techniques*, volume 1. Oxford University Press, 2004. 2.1

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. doi: http://dx.doi.org/10.1111/j.1557-9263.2009.00204.x. 2.1, 2.4, 4.7

Ryosuke O Tachibana, Naoya Oosugi, and Kazuo Okanoya. Semi-automatic classification of birdsong elements using a linear support vector machine. *PloS one*, 9(3):e92584, 2014. doi: http://dx.doi.org/10.1371/journal.pone.0092584. 2.3

Lee Ngee Tan, Kantapon Kaewtip, Martin L Cody, Charles E Taylor, and Abeer Alwan. Evaluation of a sparse representation-based classifier for bird phrase classification under limited data conditions. In *Proceedings of the Interspeech 2012 – 13th Annual Conference of the International Speech Communication Association*, pages 2522–2525, 2012. 2.3

Andrew Taylor. Recognising biological sounds using machine learning. In *AI-Conference*, pages 592–592. CiteSeer, 1995. 1.2, 2.1, 2.3

Sandra L Taylor and Katherine S Pollard. Evaluation of two methods to estimate and monitor bird populations. *PLoS One*, 3(8):e3047, 2008. doi: http://dx.doi.org/10.1371/journal.pone.0003047. 1.1

Ofer Tchernichovski. *Sound Analysis Pro 2011 User Manual*, 2012. URL `http://soundanalysispro.com/manual-1/manual-pdf`. Access date: 26/09/2016. 2.3

Ofer Tchernichovski, Fernando Nottebohm, Ching Elizabeth Ho, Bijan Pesaran, and Partha Pratim Mitra. A procedure for an automated measurement of song similarity. *Animal Behaviour*, 59(6):1167–1176, 2000. doi: http://dx.doi.org/10.1006/anbe.1999.1416. 2.2.5, 2.3

Alan James Drummond Tennyson and Paul Martinson. *Extinct birds of New Zealand*. Te Papa Press, 2006. 1.1

Frédéric E Theunissen and Sarita S Shaevitz. Auditory processing of vocal sounds in birds. *Current Opinion in Neurobiology*, 16(4):400–407, 2006. doi: http://dx.doi.org/10.1016/j.conb.2006.07.003. 2.2

Michael Towsey, Birgit Planitz, Alfredo Nantes, Jason Wimmer, and Paul Roe. A toolbox for animal call recognition. *Bioacoustics*, 21(2):107–125, 2012. doi: http://dx.doi.org/10.1080/09524622.2011.648753. 2.1, 2.2.5, 5.1

Michael Towsey, Jason Wimmer, Ian Williamson, and Paul Roe. The use of acoustic indices to determine avian species richness in audio-recordings of the environment. *Ecological Informatics*, 21:110–119, 2014. doi: http://dx.doi.org/10.1016/j.ecoinf.2013.11.007. 2.4

Vlad Trifa. A framework for bird songs detection, recognition and localization using acoustic sensor networks. *Master's thesis, École Polytechnique Fédérale de Lausanne*, 2006. 2.3, 2.2.7

Barry Truax and Gary W Barrett. Soundscape in a context of acoustic and landscape ecology. *Landscape Ecology*, 26(9):1201–1207, 2011. doi: http://dx.doi.org/10.1007/s10980-011-9644-9. 3.1

Shu-Jen Steven Tsai. Power transformer partial discharge (PD) acoustic signal detection using fiber sensors and wavelet analysis, modeling, and simulation. Master's thesis, Virginia Polytechnic Institute, 2002. 4.4.2, 4.4.3

Jari Turunen, Arja Selin, Juha T Tanttu, and Tarmo Lipping. De-noising aspects in the context of feature extraction in automated bird sound recognition. *In: Gogala, M. Trilar, T (eds.). Advances in Bioacoustics 2, Dissertationes Classis IV: Historia Naturalis, Slovenian Academy of Sciences and Arts (Ljubljana), XLVII-3*, 2006. 2.2.6, 4.4.2, 4.7, 4.7

Hemant Tyagi, Rajesh M Hegde, Hema A Murthy, and Anil Prabhakar. Automatic identification of bird calls using Spectral Ensemble Average Voice Prints. In *Proceedings of the 14th European Signal Processing Conference*, pages 1–5. IEEE, Sept. 2006. 2.3

Juan Sebastian Ulloa, Amandine Gasc, Phillipe Gaucher, Thierry Aubin, Maxime Réjou-Méchain, and Jérôme Sueur. Screening large audio datasets to determine the time and space distribution of Screaming Piha birds in a tropical forest. *Ecological Informatics*, 31:91–99, 2016. doi: http://dx.doi.org/10.1016/j.ecoinf.2015.11.012. 2.1, 2.3, 2.2.7, 2.4

Mike Unwin. *The Atlas of Birds: Diversity, Behavior, and Conservation*. Princeton University Press, 2011. 2.1

Sergey Vaisman, Shimrit Yaniv Salem, Gershon Holcberg, and Amir B Geva. Passive fetal monitoring by adaptive wavelet denoising method. *Computers in Biology and Medicine*, 42(2):171–179, 2012. doi: http://dx.doi.org/10.1016/j.compbiomed.2011.11.005. 4.4.2

P Varady. Wavelet-based adaptive denoising of phonocardiographic records. In *Proceedings of the 23rd Annual International Conference of the Engineering in Medicine and Biology Society*, volume 2, pages 1846–1849. IEEE, 2001. doi: http://dx.doi.org/10.1109/IEMBS.2001.1020582. 4.4.2

Saeed V Vaseghi. *Noise and Distortion. In: Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008. 35–50. 2.2.3, 4.3.1, 4.3.2

Lisa A Venier, Stephen B Holmes, George W Holborn, Kenneth A Mcilwrick, and Glen Brown. Evaluation of an automated recording device for monitoring forest birds. *Wildlife Society Bulletin*, 36(1):30–39, 2012. doi: http://dx.doi.org/10.1002/wsb.88. 2.4

Thiago M Ventura, Allan G de Oliveira, Todor D Ganchev, Josiel M de Figueiredo, Olaf Jahn, Marinez I Marques, and Karl-L Schuchmann. Audio parameterization with robust frame selection for improved bird identification. *Expert Systems with Applications*, 42(22):8463–8471, 2015. doi: http://dx.doi.org/10.1016/j.eswa.2015.07.002. 2.3

Beth A Vernaleo and Robert J Dooling. Relative salience of envelope and fine structure cues in zebra finch song. *The Journal of the Acoustical Society of America*, 129(5):3373–3383, 2011. doi: http://dx.doi.org/10.1121/1.3560121. 2.3

Jacques ME Vielliard. Bird community as an indicator of biodiversity: results from quantitative surveys in Brazil. *Anais da Academia Brasileira de Ciências*, 72(3):323–330, 2000. doi: http://dx.doi.org/10.1590/S0001-37652000000300006. 2.1

Erika Vilches, Ivan A Escobar, Edgar E Vallejo, and Charles E Taylor. Data mining applied to acoustic bird species recognition. In *Proceedings of the 18th IEEE International Conference on Pattern Recognition*, volume 3, pages 400–403, Aug. 2006. doi: http://dx.doi.org/10.1109/ICPR.2006.426. 2.2.6

Erika Vilches, Ivan A Escobar, Edgar E Vallejo, and Charles E Taylor. Targeting input data for acoustic bird species recognition using data mining and HMMs. In *Proceedings of the 7th IEEE International Conference on Data Mining Workshops*, pages 513–518, Oct. 2007. doi: http://dx.doi.org/10.1109/ICDMW.2007.56. 2.2.6, 2.3

TK Vintsyuk. Element-wise recognition of continuous speech composed of words from a specified dictionary. *Cybernetics and Systems Analysis*, 7(2):361–372, 1971. doi: http://dx.doi.org/10.1007/BF01071812. 2.2.6

Ciira wa Maina. Audio diarization for biodiversity monitoring. In *Proceedings of the 12th IEEE Africon International Conference - Green Innovation for African Renaissance*, pages 1–5. National Acad Sciences, Sept. 2015. doi: http://dx.doi.org/10.1109/AFRCON.2015.7331986. 2.2

David W Waite, Peter Deines, and Michael W Taylor. Gut microbiome of the critically endangered New Zealand parrot, the kakapo (*Strigops habroptilus*). *PLoS One*, 7(4):e35803, 2012. doi: http://dx.doi.org/10.1371/journal.pone.0035803. 5.2.1

Deng Wang, Duoqian Miao, and Chen Xie. Best basis-based wavelet packet entropy feature extraction and hierarchical EEG classification for epileptic detection. *Expert Systems with Applications*, 38(11):14314–14320, 2011. doi: http://dx.doi.org/10.1016/j.eswa.2011.05.096. 4.4.1

Peter M Waser and Mary S Waser. Experimental studies of primate vocalization: Specializations for long-distance propagation. *Zeitschrift für Tierpsychologie*, 43(3):239–263, 1977. doi: http://dx.doi.org/10.1111/j.1439-0310.1977.tb00073.x. 3.1

Robert Wielgat, Tomasz Potempa, Pawel Święt ojański, and Daniel Król. On using prefiltration in HMM-based bird species recognition. In *Proceedings of the IEEE International Conference on Signals and Electronic Systems*, pages 1–5, Sept. 2012. doi: http://dx.doi.org/10.1109/ICSES.2012.6382258. 2.3

Linda Wilbrecht and Fernando Nottebohm. Vocal learning in birds and humans. *Mental Retardation and Developmental Disabilities Research Reviews*, 9(3):135–148, 2003. doi: http://dx.doi.org/10.1002/mrdd.10073. 2.2

Wildlife Acoustics, Inc. *Song Scope Bioacoustics Software Version 4.0 Documentation*, 2011. URL `http://www.wildlifeacoustics.com/images/documentation/Song-Scope-Users-Manual.pdf`. Access date: 26/09/2016. 2.2.1, 2.3

R Haven Wiley and Douglas G Richards. Physical constraints on acoustic communication in the atmosphere: Implications for the evolution of animal vocalizations. *Behavioral Ecology and Sociobiology*, 3(1):69–94, 1978. doi: http://dx.doi.org/10.1007/BF00300047. 3.1, 3.4.1, 3.4.1, 3.4.2

E Williams. Australasian bittern : *In* Miskelly, C.M. (ed.) New Zealand Birds Online, 2013. URL `www.nzbirdsonline.org.nz`. Access date: 04/07/2016. 5.1, 5.2.1

Heather Williams. Birdsong and singing behavior. *Annals of the New York Academy of Sciences*, 1016(1):1–30, 2004. doi: http://dx.doi.org/10.1196/annals.1298.029. 7

Jason Wimmer, Michael Towsey, Paul Roe, and Ian Williamson. Sampling environmental acoustic recordings to determine bird species richness. *Ecological Applications*, 23(6):1419–1428, 2013. doi: http://dx.doi.org/10.1890/12-2088.1. 1.2, 2.1

Katie Wolf. Bird song recognition through spectrogram processing and labelling, 2009. 2.2.3, 2.1, 5.1

Zunjing Wu and Zhigang Cao. Improved MFCC-based feature for robust speaker identification. *Tsinghua Science & Technology*, 10(2):158–161, 2005. doi: http://dx.doi.org/10.1016/S1007-0214(05)70048-1. 2.2.6

Marius Zbancioc and Mihaela Costin. Using neural networks and LPCC to improve speech recognition. In *Proceedings of the IEEE International Symposium on Signals, Circuits and Systems*, volume 2, pages 445–448, July 2003. doi: http://dx.doi.org/10.1109/SCS.2003.1227085. 2.2.6

Xiaoxia Zhang and Ying Li. Adaptive energy detection for bird sound detection in complex environments. *Neurocomputing*, 155:108–116, 2015. doi: http://dx.doi.org/10.1016/j.neucom.2014.12.042. 2.2, 2.2.6, 2.2.7, 2.2.7

Sue Anne Zollinger and Henrik Brumm. Why birds sing loud songs and why they sometimes don't. *Animal Behaviour*, 105:289–295, 2015. doi: http://dx.doi.org/10.1016/j.anbehav.2015.03.030. 2.2

Mieke C Zwart, Andrew Baker, Philip JK McGowan, and Mark J Whittingham. The use of automated bioacoustic recorders to replace human wildlife surveys: An example using nightjars. *PloS one*, 9(7):e102770, 2014. doi: http://dx.doi.org/10.1371/journal.pone.0102770. 2.4

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION TO DOCTORAL THESIS CONTAINING PUBLICATIONS

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of Candidate: Nirosha Priyadarshani

Name/Title of Principal Supervisor: Stephen Marsland

Name of Published Research Output and full reference:

Birdsong Denoising Using Wavelets

Priyadarshani N, Marsland S, Castro I, Punchihewa A (2016) Birdsong Denoising Using Wavelets. PLoS ONE 11(1): e0146790. doi:10.1371/journal.pone.0146790

In which Chapter is the Published Work: Chapter 4

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate:

  and / or

- Describe the contribution that the candidate has made to the Published Work:

  Conceived and designed the experiments: SM NP IC. Performed the experiments: NP. Analyzed the data: NP SM IC AP. Contributed reagents/materials/analysis tools: NP IC. Wrote the paper: NP IC SM AP.

Nirosha Priyadarshani

Digitally signed by Nirosha Priyadarshani
DN: cn=Nirosha Priyadarshani, o=Massey University, ou=SEAT, email=N.Priyadarshani@massey.ac.nz, c=NZ
Date: 2016.12.14 15:14:32 +13'00'

Candidate's Signature _____ Date _____

Digitally signed by Stephen Marsland
DN: cn=Stephen Marsland, o=Massey University, ou=SEAT, email=s.r.marsland@massey.ac.nz, c=NZ
Date: 2016.12.15 11:29:33 +13'00'

Principal Supervisor's signature _____ Date _____