Digital Signal Processing Approaches to Bird Song Analysis

A dissertation submitted to the University of Dublin
for the degree of Doctor of Philosophy

Colm O'Reilly
Trinity College Dublin, October 2017

Signal Processing and Media Applications
Department of Electronic and Electrical Engineering
Trinity College Dublin



To my family and friends.

Abstract

The ability to automatically analyze bird vocalizations would provide major assistance to zoologists in their behavioral and ecological studies. Revisions of taxonomy need to be made in cases where a population of species has evolved or diversified enough to be treated as a new unique species. Bird vocalizations are an important facet of the review process as song has a major influence in mate choice, hence why zoologists tend to study distinguishable songs when analyzing populations.

This thesis presents a method to quantify differences in bird vocalizations, inspired by dialect difference assessment from speech processing. The use of different codebooks is introduced. Reference to how easily pairs can be separated using Gaussian Mixture Model (GMM) classifiers with Mel-frequency Cepstral Coefficients (MFCCs) is also presented. As endangered populations will inevitably have small datasets, machine learning approaches may be hard to train. A forensics inspired pitch contour difference measure works with approximately 200 samples per population.

An improved pitch tracker for bird vocalizations, called YIN-bird, is also introduced. Carefully tuned parameters improve pitch tracking with YIN (a commonly used pitch tracking tool in the speech community), but optimal parameters can change quickly even within one song. YIN-bird, a modified version of YIN which exploits spectrogram properties to automatically set a minimum Fundamental Frequency ($\boldsymbol{F_0}$) parameter for YIN, is presented. Gross pitch errors on whistles and trills are reduced by up to 4% on a ground truth dataset of synthetic bird song with known pitch which is an improvement. This dataset was evaluated by expert listeners and described as "sounding like original & can hardly tell it is synthetic". A qualitative analysis of complex bird vocalizations is also presented.

Work then focuses on techniques to automatically extract acoustic features commonly used by zoologists. Signal processing techniques are employed to automate the extraction of the acoustic features: maximum, minimum and peak frequency, and bandwidth. YIN-bird and sine-tracking, a feature extraction method successfully applied to bird classification previously, are the automatic methods employed. The performance of automatic methods is compared to the manual method currently used by zoologists. Results are compared to a major study on species delimitation in the zoology domain. Both methods are well suited to this task, and demonstrate the strong potential to begin to automate the task of acoustic comparison of bird species.

The thesis then addresses the automatic segmentation of bird recordings into target and non-target regions using a speaker diarization approach. Standard approaches to automatic bird species identification are inspired by speaker verification methods, and tend to use all available bird vocalizations in a sample. Zoologists however will first tend to isolate important parts of a vocalization, discarding less discerning elements. Thus the automation of such segmentation work is highly desirable. The first stage is to segment target vocalizations from field recordings which is performed automatically in this thesis. Birds are typically recorded in their natural environment with many competing signals present. A speaker diarization approach to segmenting such recordings into target/non-target events is presented. This diarization approach uses Bayesian Information Criterion (BIC) Hierarchical Agglomerative Clustering (HAC) from the LIUM toolkit. MFCCs are used as features. The classification error is reduced from 19.20% to 15.01% when comparing the proposed system to a HMM-based segmentation as a baseline. A discussion of

system performance and conditions which influence false alarms is given.

Declaration

I hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other University and that it is entirely my own work.

I agree that the Library may lend or copy this thesis upon request.

Signed,

Colm O'Reilly

October 9, 2017.

Acknowledgments

First off I would like to thank my supervisor Dr. Naomi Harte for her continuous support during my Ph.D study, for her patience, motivation, and immense knowledge. I could not have imagined a better advisor and mentor for my Ph.D study.

I would also like to thank Prof. Boland, Dr. Corrigan, Dr. Pitié, Dr. Hines and Prof. Kokaram for their encouragement during my research. A special thank you to Ailbhe, Finnian, Eoin and Ian who always had time for my questions when I was starting my research and to all the students I have worked with since. I wish the best of luck to Matt and George, my Ph.D colleagues. I would also like to thank all the staff who made my time at TCD so enjoyable, especially John for solving all my computer problems over the years, the man is never in bad form.

I would like to express gratitude to Prof. Alwan and Prof. Taylor at UCLA, who welcomed me for a semester during my studies. Thank you to Kantapon (Jom), it was a pleasure working with you I look forward to collaborating again in the future. Thank you to everyone I crossed paths with at UCD, especially Prof. Sheridan for his compelling advice. To my UCD housemate and cousin Conor, I was honored to live those days with you.

A Ph.D can be stressful and melancholy, but luckily I have a great bunch of friends whose humor and entertainment helped me unwind on weekends. Thanks to the Roadies (you know who you are) whose banter filled pointless Whatsapp conversations offered much needed light relief during long hours in the lab. A special mention for my housemates and golf partners who offered retreat from the lab during my studies.

None of this would be possible without the love and support of my parents. Their constant encouragement, praise and trust in me is incredible. Observing their work ethic first hand growing up gave me the confidence to be ambitious. Thank you to Ade and PJ who have always felt like a second set of parents to me. Thank you to aunty Maureen, it's wonderful to have you in Roscommon. Thank you to uncle Pio, to many more days out in Croker, Sam will reach Mayo very soon. A special thanks also to Laverne, Breen and Pat.

And of course last but by no means least a big thank you to my beautiful girlfriend Crona. Your support and patience over the last few years means the world to me. You put up with extended hours of neglect without complaint while I worked long nights on this thesis and I promise to make up every minute to you. Here's to the future!

Contents

Contents vii

List of Acronyms xi

1 Introduction 1
- 1.1 Thesis outline . . . 2
- 1.2 Publications . . . 3

2 A review of bird acoustics analysis 7
- 2.1 Bird Vocalizations . . . 8
  - 2.1.1 Bird Communication . . . 8
  - 2.1.2 Influences on vocal learning . . . 8
  - 2.1.3 Production and Perception . . . 9
- 2.2 Why is song important? . . . 10
- 2.3 Characterization of vocalizations . . . 12
- 2.4 Automatic acoustic analysis by engineers . . . 16
- 2.5 Bird species classification . . . 17
- 2.6 Detection of bird events . . . 21
  - 2.6.1 Presence/absence . . . 22
  - 2.6.2 Voice Activity Detection (VAD) . . . 22
  - 2.6.3 Polyphonic Segmentation . . . 24
  - 2.6.4 Time-frequency Boxes . . . 24
  - 2.6.5 Time-frequency Blobs . . . 25
  - 2.6.6 Time-frequency Sinusoids . . . 25
  - 2.6.7 Diarization applied to bioacoustics . . . 25
- 2.7 Bird Taxonomy . . . 27
- 2.8 Quantifying difference of bird vocalizations . . . 28
- 2.9 System Implementation . . . 31
  - 2.9.1 Mel-frequency Cepstral Coefficients (MFCCs) . . . 31
  - 2.9.2 Classification using Gaussian Mixture Models (GMMs) . . . 32
  - 2.9.3 Testing phase of GMM classifiers . . . 32

2.9.4 Classification using Hidden Markov Models (HMMs) . . . 33
2.10 Collaboration of Engineering and Zoology . . . 34

3 Quantifying difference in vocalizations of bird populations 37
3.1 Method - Spoken dialect pitch contour difference . . . 38
3.2 Bird Data . . . 41
3.2.1 Indonesian island populations - Set 1 & 2 . . . 41
3.2.2 Species - Alder Flycatcher & Barn Swallow - Set 3 . . . 43
3.2.3 Mexican antbirds - Set 4 . . . 43
3.2.4 GMM Classification . . . 44
3.3 Dialect difference for bird vocalizations . . . 46
3.3.1 Difference measure with 3x3 codebook . . . 46
3.3.2 Extension to a 5x5 codebook . . . 48
3.4 Accuracy of pitch contours . . . 49
3.5 Comparison of F_0 estimation using YIN-bird for bird call difference . . . 50
3.6 Bird call difference with YIN-bird . . . 57
3.6.1 YIN-bird setup for bird difference measure . . . 57
3.6.2 3x3 codebook difference measure with YIN-bird . . . 57
3.6.3 5x5 Codebook difference measure with YIN-bird . . . 59
3.7 Vector Quantization (VQ) Codebook . . . 60
3.7.1 VQ codebook per pair - 9 Centers . . . 60
3.7.2 VQ codebook per pair - 25 Centers . . . 62
3.8 Data size and self comparison . . . 64
3.8.1 3x3 & 5x5 codebook with reduced data size and self comparison . . . 64
3.8.2 VQ codebook with reduced data size and self comparison . . . 64
3.9 Discussion . . . 67

4 YIN-bird: Improved pitch tracking for birds 73
4.1 Pitch Tracking . . . 73
4.1.1 Pitch Importance . . . 74
4.1.2 Pitch Tracking Techniques . . . 74
4.1.3 Current pitch tracking performance . . . 76
4.1.4 Pitch Tracking for Birds . . . 77
4.2 Tuning YIN to Bird Song . . . 79
4.3 Experimental Setup . . . 83
4.3.1 Data . . . 83
4.3.2 Synthesized Bird Sounds . . . 83
4.3.3 Listener Tests . . . 85
4.3.4 Error Metrics . . . 86

4.3.5 Experiment Parameters . . . 86
4.4 Quantitative Results . . . 87
4.5 Qualitative Analysis . . . 93
4.5.1 Harmonic (or Nasal) pitch tracking . . . 93
4.5.2 Two-voiced . . . 94
4.5.3 Noise . . . 94
4.5.4 Biphonation . . . 97
4.5.5 Birds with complex vocalizations . . . 98
4.5.6 Signal Degradation . . . 101
4.6 Discussion . . . 101

5 Species delimitation inspired by zoological techniques 103
5.1 What parts of vocalizations do Zoologists see as important? . . . 104
5.1.1 Species Delimitaion . . . 104
5.1.2 Tobias et al. Experiments . . . 105
5.1.3 Tobias et al. strengths and limitations of system . . . 108
5.1.4 Acoustic Evidence . . . 108
5.2 Automatically Extracting Features . . . 111
5.2.1 Temporal Features . . . 111
5.2.2 Feature Extraction using YIN-bird . . . 111
5.2.3 Feature Extraction using Sine Wave Tracking . . . 112
5.3 Experiments . . . 112
5.3.1 Data . . . 112
5.3.2 Expert Labeling of Data using *Praat* . . . 113
5.3.3 Results . . . 118
5.3.4 Discussion . . . 125
5.4 Comparison: Bird difference measure vs. Tobias measure . . . 125
5.4.1 Why difference measures don't always agree . . . 127
5.5 Conclusion . . . 128

6 Segmentation of Field Recordings 133
6.1 Segmentation of Bird Song . . . 134
6.1.1 Speaker Diarization . . . 135
6.2 Experimental Setup . . . 136
6.2.1 Data for Segmentation . . . 137
6.2.2 Evaluation . . . 139
6.2.3 Segmentation using Hidden Markov Models (HMMs) . . . 141
6.2.4 Segmentation using speaker diarization tool LIUM . . . 142
6.3 Results . . . 144

6.4 Discussion 146
6.4.1 Qualitative discussion of segmentation 148
6.4.2 Extension of Diarization and Clustering 150
6.5 Conclusion 153

7 Conclusion **155**
7.1 Chapter 2: A review of bird acoustic analysis 155
7.2 Chapter 3: Quantifying difference in vocalizations of bird populations 155
7.3 Chapter 4: YIN-bird: Improved pitch tracking for birds 156
7.4 Chapter 5: Species delimitation inspired by zoological techniques 156
7.5 Chapter 6: Segmentation of Field Recordings 157
7.6 Final Remark 157

A Species delimitation data **159**

B Hidden Markov Models (HMMs) theory **167**

Bibliography **173**

List of Acronyms

AC Autocorrelation

AF Alder Flycatcher

BIC Bayesian Information Criterion

BNO Black-naped Oriole

BOU British Ornithologists' Union

BS Barn Swallow

BU Buton

CEP Cepstrum

CLR Cross Likelihood Ratio

CNN Convolutional Neural Network

DAB Dusky Antbird

DCT Discrete Cosine Transform

DTW Dynamic Time Warping

DWA Dot-winged Antwren

EER Equal Error Rate

EM Expectation Maximization

F_0 Fundamental Frequency

FFT Fast Fourier Transform

F_s Sampling Rate

GLR Generalized Likelihood Ratio

GMM Gaussian Mixture Model

HAC Hierarchical Agglomerative Clustering

HMM Hidden Markov Model

HPS Harmonic Product Spectrum

IFFT Inverse Fast Fourier Transform

KL Kullback Leibler

LL Log-likelihood

LNS Library of Natural Sounds, Macaulay, University of Cornell

LPC Linear Prediction Coding

MAT Mexican Anthrush

MFCC Mel-frequency Cepstral Coefficient

ML Maximun Likelihood

MTM Medial Tympaniform Membrane

NaN Not a Number

NCCF Normalized Cross-correlation Function

NMF Non-negative Matrix Factorization

NSA National Sound Archive, British Museum

OBSB Olive-backed Sunbird

PCN percentage of correct non-speech identification

PLP Perceptual Linear Prediction

PCS percentage of correct speech identification

PDA Pitch Detection Algorithm

PF percentage of false identification

SAP Sound Analysis Pro

SHR Subharmonic-to-harmonic Ratio

SHS Subharmonic Summation

SMS Spectral Modeling Synthesis

SNR Signal-to-Noise Ratio

SpR Sine plus Residual

SRH Summation of Residual Harmonics

STFT Short Time Fourier Transform

SU Sulawesi

SVM Support Vector Machine

TCD Trinity College Dublin

UBM Universal Background Model

VAD Voice Activity Detection

VQ Vector Quantization

WW Wangi-Wangi

XC Xeno-canto.org

1 Introduction

The field of zoology is constantly studying bird populations to measure species abundance and review species limits for conservation purposes. These conservation efforts benefit from accurate automatic identification of bird species. Knowing what species populate a given area currently requires a lot of manual effort from researchers in the form of listening to hours of data. Automatic monitoring would reduce this effort considerably. Another issue in ornithology is determining how different vocalizations of bird populations are, directly linked to taxonomic classification of populations. Reclassification of bird populations requires analysis of various traits such as biometric measurements, mating rituals and acoustic information. Digital signal processing and machine learning techniques explored in this thesis help automate analysis of acoustic evidence. Other work in the field includes measuring variability within a species song repertoire, classifying phrases, tracking vocal learning and measuring species abundance automatically. Commercial applications, which allow bird enthusiasts to identify birds in their local park using their phones, are another possible avenue for bird song analysis.

There is growing interest in the study of bird song within the speech processing community. Bird species classification systems have been adapted from speaker identification technologies. When zoologists analyze bird populations, they tend to look at acoustic evidence in specific ways. This thesis investigates "Digital Signal Processing Approaches to Bird Song Analysis". Work here helps engineers and zoologists automate their acoustic analysis of bird song. Work in this thesis investigates vocal difference between bird populations using improved pitch tracking techniques. Automatic segmentation of field recordings is also explored.

1.1 Thesis outline

This thesis is outlined as follows.

Chapter 2: A review of bird song analysis

This chapter provides background on bird song analysis used in subsequent Chapters. An overview of bird vocalizations is given and their importance stated. Characteristics of vocalizations are described in detail. Automatic acoustic analysis by engineers is reviewed. Previous work on bird species classification are examined and relevant techniques described. Next the detection of bird acoustic events is discussed. This work provides background to segmentation work presented in chapter 6. Work in chapter 6 uses a novel speaker diarization method which is introduced in chapter 2. A brief description of bird taxonomy is also given which provides information of why zoologists wish to analyze vocal difference between populations. Methods used in this thesis which quantify bird vocalization differences are introduced in Section 2.8. A technical overview of system implementations such as Mel-frequency Cepstral Coefficients (MFCCs), Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) is also given. Finally a discussion on how work in this thesis merges work from the fields of engineering and zoology to aid researchers with their work is provided.

Chapter 3: Quantifying difference in vocalizations of bird populations

Work in chapter 3 takes a dialect difference system first developed for speech and refines it to measure vocalization difference between bird populations by extracting pitch contours. The pitch contours are transposed into pitch codes. A variety of codebook schemes are discussed. A novel Vector Quantization (VQ) approach to representing the data is implemented. Datasets are introduced and results presented.

Chapter 4: YIN-bird: Improved pitch tracking for birds

This chapter presents YIN-bird, an optimized Pitch Detection Algorithm (PDA) for bird vocalizations. During development of work in chapter 3, it became clear that PDAs that work well for speech did not work as well on bird song. The optimization of YIN (a well known PDA used in speech processing) to develop YIN-bird is presented. An evaluation of how YIN-bird performs on various bird syllable types is given quantitatively. A number of complex syllable types are evaluated qualitatively. Synthetic data was developed to evaluate YIN-bird and listener tests assessed the quality of the data. It is the first known study to develop a dataset of synthetic bird song to allow comparison of pitch extraction for bird song with a ground truth. The dataset has been publicly shared. This work represents significant progress in understanding the challenges of pitch extraction of bird song. YIN-bird is used in chapter 3 (for final experiments) and chapter 5 (throughout).

Chapter 5: Species delimitation inspired by zoological techniques

This chapter deals with species limits using the system from [160]. Joseph Tobias is a zoologist who proposed a points-based system which measures morphology, acoustic, plumage, behavioral and geographical evidence to determine species status of bird populations. Technical work in this chapter presents the use of YIN-bird and Sine Tracking [65] to automatically extract acoustic features. Work in [160] required manual analysis of spectrograms using on-screen cursors. Chapter 5 presents automatic means of acoustic feature extraction and evaluates its performance. This work takes its lead from the treatment of acoustic comparison of birds directly from the zoology community and seeks to automate bird song comparison in a manner that has the potential to become accepted in the zoology community. The use of data used in the zoology study strengthens this contribution further.

Chapter 6: Segmentation of Field Recordings

A recurring theme in this work has been the isolation of segments of song from longer field recordings. The expertise built up in the speaker diarization community is brought to the domain of bird recordings in this Chapter. The work presented demonstrates the strong potential for methods based on Hierarchical Agglomerative Clustering (HAC) to accurately segment bird song. The majority of acoustic bird data is recorded in the bird's natural habitat which can be corrupted by environmental noise and non-target vocalizations. This chapter presents a speaker diarization approach to automatically segmenting field recordings. Currently datasets are manually annotated by expert listeners which is laborious and subjective. The speaker diarization approach is compared with baseline Hidden Markov Model (HMM) system trained on non-target background and target song data.

Chapter 7: Conclusion

This chapter draws together the main conclusions of this thesis, and suggests possibilities of future work.

1.2 Publications

Portions of the work described in this thesis have appeared in the following publications:

Conference Papers

- Colm O'Reilly, Nicola M. Marples, David J. Kelly, and Naomi Harte, "YIN-Bird: Improved Pitch Tracking for Bird Vocalisations", in *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association*, September 8–12, San Francisco, California, USA, pages 2641–2645.

- Colm O'Reilly, Nicola M. Marples, David J. Kelly, and Naomi Harte, "Quantifying difference in vocalizations of bird populations", in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, September 6–10, Dresden, Germany, pages 3417–3421.

Journal Articles

- C. O'Reilly and N. Harte, "Pitch tracking of bird vocalizations and an automated process using YIN-bird", in Journal of Cogent Biology, 2017. (Accepted with minor revisions)

- K. Kaewtip, A. Alwan, C. O'Reilly, and C. E. Taylor, "A robust automatic birdsong phrase classification: a template-based approach", in the Journal of the Acoustical Society of America 2016.

Conference Papers under review

- C. O'Reilly, C. Saam and N. Harte, "Automatic Segmentation of Bird Song Recordings Using Speaker Diarization", in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*, August 20-24, Stockholm, Sweden, 2017. (Under review)

- C. O'Reilly, M. Köküer, P. Jančovič, R. Drennan and N.Harte, "Automatic Frequency Feature Extraction for Bird Species Delimitation" in 25th European Signal Processing Conference (EUSICPO 2017), Aug 28 - Sept 2, Kos, Greece, 2017. (Under review)

Presentations

- Poster presentation: C. O'Reilly, N. M. Marples, D. J. Kelly, and N. Harte, "YIN-bird: Improved Pitch Tracking for Bird Vocalisations", in INTERSPEECH 2016, September 9-12, San Francisco, USA.

- Poster presentation: C. O'Reilly, N. M. Marples, D. J. Kelly, and N. Harte, "YIN-bird: Improved Pitch Tracking for Bird Vocalisations", in UK Speech 2016, University of Sheffield, UK.

- Poster presentation: C. O'Reilly, N. M. Marples, D. J. Kelly, and N. Harte, "Quantifying difference in vocalizations of bird populations", in Into the Wild 2015, Queen Mary University of London, UK.

- Poster presentation: C. O'Reilly, N. M. Marples, D. J. Kelly, and N. Harte, "Front-end Filters for Bird Call Feature Extraction", in UK Speech 2014, University of Edinburgh, UK.

- Talk: “YIN-bird: Improved Pitch Tracking for Bird Vocalisations”, Speech Communication Lab groups seminar, TCD, Ireland, 2016.
- Talk: “Quantifying difference in vocalizations of bird populations”, Doctoral Workshop and INTERSPEECH 2015, September 6-10, Dresden, Germany, 2015.
- Talk: “Species Similarity”, Taylor Labs, UCLA, USA, 2015.
- Talk: “Dialect Distance for Birds”, SPAPL, UCLA, USA, 2015.
- Talk: “Species Recognition”, Sigmedia seminar, TCD, Ireland, 2014.

2

A review of bird acoustics analysis

The ability to automatically analyze bird vocalizations could provide major assistance to zoologists in their behavioral and ecological studies. The importance of birds' vocalizations in these studies cannot be overstated. Bird vocalizations are generally described as calls (short and monosyllabic) or song (long complex and melodic) [25]. Bird song is essential for communication, especially for mate attraction and territory defense [25]. When visibility is limited, such as in rainforests with dense vegetation, acoustic communication may be the only means of species identification [165]. Bird monitoring is often performed by acoustic means as collection of bird sounds is more convenient than vision or other indicators [148]. In addition, even for species for which visually based methods are feasible, passive acoustic methods offer the potential for greater detection ranges in some environments (e.g. in dense forest), and hence potentially better precision [97].

While the scientific study of bird song has made important contributions to the field of zoology, its intrigue has also sparked interest from speech and language researchers in an effort to improve the efficiency, accuracy, and repeatability of bird song analysis to monitor and assess bird communities [38]. This chapter will consist of a brief introduction to acoustic analysis of bird vocalizations. The importance, influence of learning, production, perception, song development and characteristics of bird vocalizations from a biodiversity context will be discussed. Automation of acoustic analysis by engineers will then be reviewed. Techniques used by zoologists to set species limits will be examined, a process which holds acoustic evidence in high regard. Finally a discussion follows of how work in this thesis contributes to the collaborative research of bird song, using ideas from both zoology and engineering.

2.1 Bird Vocalizations

A brief overview of how birds use vocalizations to communicate, influence on vocal learning and production and perception of vocalizations will now be given.

2.1.1 Bird Communication

"Bird song: biological themes and variations" by Catchpole and Slater [25] gives a comprehensive guide to bird song for those who are not zoological experts. The book contains many interesting topics one of which is bird communication. Birds sing to communicate with other members of the same species. Birds can easily identify conspecific[1] or heterospecific[2] neighbors from their calls and songs [32]. If the behavior of a receiving animal changes then communication has taken place. Calls between the receiver and sender can be cooperative or as part of conflict. Species' hearing evolves to become sensitive to their populations calls [7]. Birds sometimes use cost minimizing conspiratorial whispers to conserve energy. In contrast the male tit sings for as long and as loud as he can manage simply to force his message across [25]. When males are defending their territorial boundaries, loudness and repetition are a strong feature of their vocalizations. Females often reject weak mating calls and males who sing at a faster rate attract birds quickly. Information which is generally communicated by calls are species, sex, identity and likely future actions. Birds are usually good at detecting false information and it can be quite costly for a smaller, weaker bird to cheat or bluff the receiver into accepting it as a larger, stronger rival or mate [25]. Vocalizations are the most effective form of bird communication. They can be effective day or night and require less energy than movement. Song is particularly useful in environments like the canopy of a rainforest where visibility is impaired [25]. The effectiveness of vocalizations in species recognition by birds themselves, acts as strong indication that scientists can learn a lot about populations from acoustic data.

2.1.2 Influences on vocal learning

Birds mainly learn from their species neighbors which is why vocal learning will differ between individuals living in the wild, living in isolation and those who are deafened [96]. Catchpole and Slater's book describes how nature and nurture affect song. Birds raised in isolation will not have the same repertoire as birds reared among their own species. Vocal learning usually takes place within the first year. If a bird is tutored after its first year and is in full song it will not alter its song. Birds learn more in their youth and their sensitivity to song exposure decreases as age increases. Bird populations can alter their song year to year to fit in with groups. Two subspecies will differ in song from one another. But if young males from subspecies A are kept in the company of females from subspecies B, they develop the song of that subspecies rather than their own [25].

[1] Conspecific: Different species
[2] Heterospecific: Same species

Learning song is important for the evolution of a species. Learning is a good mechanism for the transmission of complex information. Learning allows social adaptation, avoids conflict and increases the breeding success when surrounded by neighbors. Learning also provides an honest indicator of male quality [25]. Birds learn song from tutors. Tutors can be their father or territorial neighbors. Aggressive tutors are the best. Imitation is common among birds. Migrant bird populations have to learn songs quickly or the species will die out.

A number of other factors can influence a bird's vocalizations. The stage of the species' breeding cycle affects the song. The presence of neighboring subjects and predators can force a bird to alter its calls and songs. Environmental conditions such as time of day, time of year, weather conditions, vegetation and background noise also influences choice of song. The position of the speaker and observer is another key influence [25].

2.1.3 Production and Perception

Like in humans, bird sounds are produced by the flow of air during expiration through a vocal system [46]. Bird vocalization production, like speech production, can be modeled by a source-filter mechanism. Some bird vocalizations are tonal (e.g. Sparrow song), and some have a noisy spectrum quality where multiple stacked frequency components are present like humans (e.g. Zebra Finchs) [46]. It had been suggested that because pure-tone song is common in birds, perhaps their vocal system is more like a whistle. This hypothesis was rejected by Beckers et al. [8], where direct evidence of a source-filter mechanism in the production of pure-tone bird song was presented. Beckers et al. presented more evidence in [7], which suggests bird song production and perception are very similar to humans. Human voiced speech is made up of strong energy at Fundamental Frequency ($\boldsymbol{F_0}$) with relative amplitudes at multiple harmonics due to the properties of humans' source filter system. Bird vocalizations have higher pitch than humans which means the interval between harmonics is larger than for human speech. This implies that even gentle low pass filtering by a bird's vocal tract could potentially remove all harmonics and leave just $\boldsymbol{F_0}$. This is the case with tonal (single tone) sound production. In contrast, overtone or nasal sounds are produced when the bird uses a wider bandwidth filter which allows sounds at $\boldsymbol{F_0}$ and multiple harmonics to be emitted during vocalization.

Birds produce sound using their equivalent of the human voice box called the syrinx. Whereas the human larynx is situated at the top of the trachea, the syrinx is much lower down, at the junction of the two bronchi. This means that the syrinx has two potential sound sources (voices), one in each bronchus (see Figure 2.1 [49]). The sounds are mixed when fed into the common trachea and buccal cavity [25]. One side of the syrinx usually dominates the other source at any particular time. Some syllables are produced from the right and some from the left. Some syllables are a combination of both. The vocal tract modulates the signal in the same way a human's throat would but there is a greater distance from the syrinx to the point where the vocalizations enter the outside world (relative to the size of the subject) [25].

The syrinx contains membranes with the most important being the Medial Tympaniform Membranes (MTMs) situated in the medial wall of the bronchi. Sound is produced by the vibrations of MTM against the cartilage wall [49]. If the MTM is punctured in a male non-songbird, he can no longer sing. This is not the case for songbirds [25]. The larynx in birds does not include vocal folds like in humans. The larynx plays little to no role in sound production but together with the mouth and beak acts as a resonator to sound produced by the syrinx [49].

2.2 Why is song important?

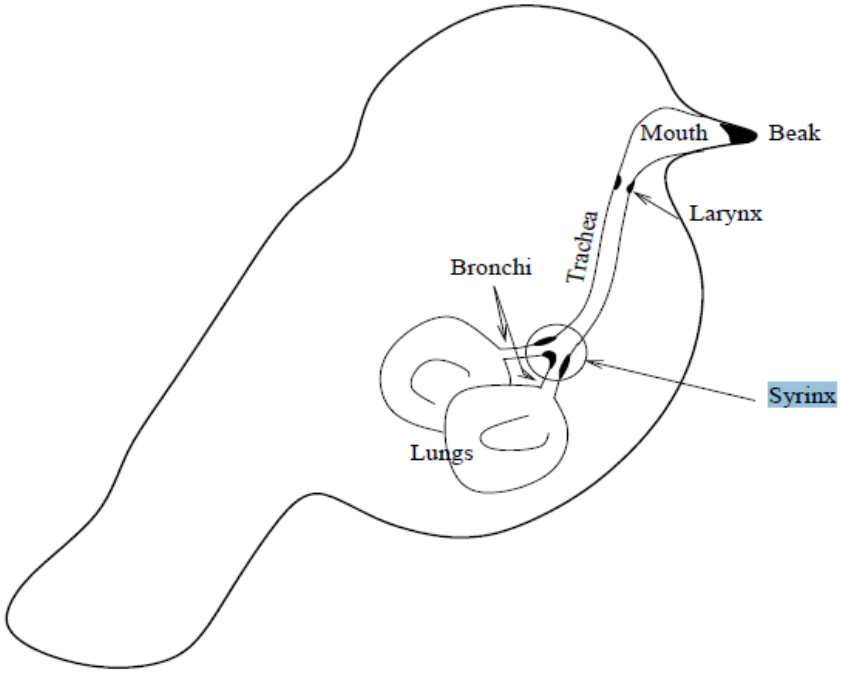
Conservation concerns have motivated the increased study of the geographic distribution of bird species. A populations' vocalization, particularly their song, can be used to identify species. Classification of animals by their sounds is valuable for biological research and environmental monitoring applications [88]. In addition to this, many amateurs and enthusiasts undertake similar activities as a hobby [145]. Identifying call variation can give interesting insight into geographical patterns within a given bird species. As most bird vocalizations have evolved to be species-specific, it is a natural and adequate way to automatically identify species [88].

Populations of a bird species can evolve over time to become new species. While plumage patterns and other morphological information can remain constant, the vocalizations of a given population may have diversified enough to warrant reclassification. There are a number of reasons why acoustic analysis of avian taxonomy has increased over the past few years [63, 3].

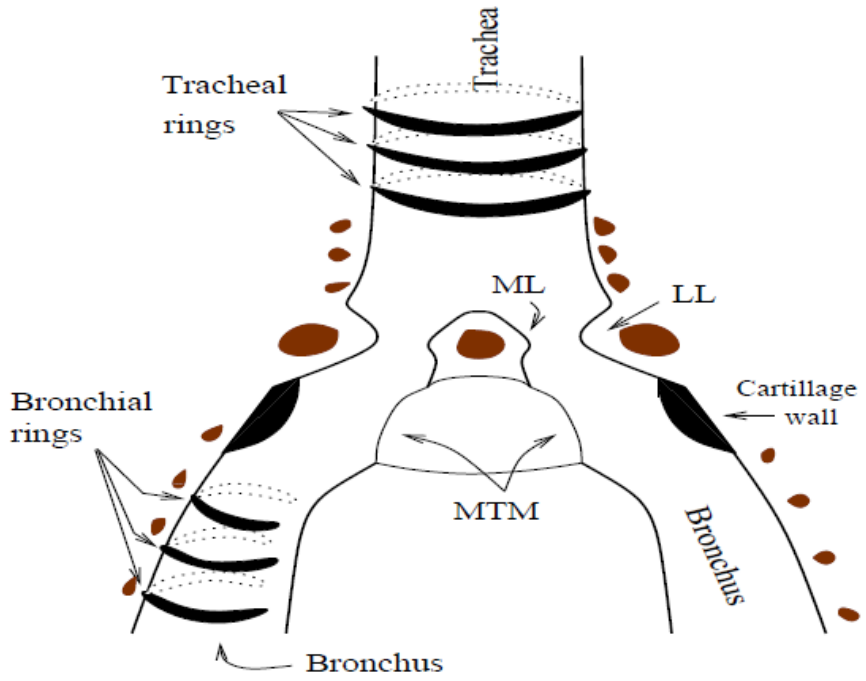
- Firstly, acoustic signals often play a central role in species recognition and mate choice, and therefore mediate reproductive isolation in many avian systems [160]. Catchpole et al. in [25], emphasized vocalization importance in mate choice and species recognition which suggests acoustic signals may give early clues of species distinction [86]. Not only can acoustic divergence precede divergence of other evidence, some even argue bird vocalizations can sometimes provide a better indication of species limits than morphological characters [98, 119, 132, 160].

- Secondly, the use of songs and calls to delimit species and monitor populations have several practical advantages, e.g. ease and economy of sound recording and analysis [130]. In difficult environments, like the canopy of rainforests, recording birds singing is more straightforward than capturing individual birds and taking morphological measurements. Birds are actually not easy to photograph. Birds are most often hidden, perched high in a tree or frightened by human presence. They can also fly very quickly [69]. Taking photographs of birds under these circumstances is quite difficult, whereas audio recordings of calls and songs have proven easier to collect [69]. Using audio records rather than bird pictures is justified by current practices [19, 163, 165].

- Thirdly, collections of sound libraries like xeno-canto.org [166], the Cornell Laboratory of Ornithology, Macaulay Library [118] and the British Sound Archive [90] now rival museum



(a)



(b)

Figure 2.1: Bird vocal system. (a) Syrinx shown in relation to a bird's anatomy [49]. (b) Schematic view of bird syrinx (bird's equivalent to the human larynx). Sound is produced by the vibrations of MTM against the cartilage wall [49]. Soft tissues medial and lateral labia (ML and LL) help control the MTM vibrations.

specimen collections and online genetic databases in the diversity of taxa represented and the depth of sampling per taxon [160]. For these reasons ornithologists are interested in multiple applications such as classification, feature extraction, isolation of segments of song, and measurement of call and song difference in birds. Currently many of these applications are performed under manual supervision but given the expansion of available data, researchers require automatic methods to deal with ever increasing libraries of bird recordings.

2.3 Characterization of vocalizations

Birds produce a wide variety of vocalizations. These range from short, monosyllabic calls, to long complex song [25]. Early researchers did not agree on a common set of units by which bird song of various different species might be described [159]. Thompson et al. [159] presented a system for describing bird song units in the hope of greater standardization in the protocols by which researchers generate and name bird song units. A note or element refers to the smallest level of song (which can be analogous to phonetic units) and is defined by a sound represented by a continuous trace on the spectrogram. Notes can be grouped together to form syllables, which are units of sound separated by silent intervals [46]. Syllables and notes are themselves organized into third order units known as phrases, and phrases are in turn clustered together into performances called song [25, 159]. Labeling units of bird song still differs from scientist to scientist, and species to species, but an example of unit segmentation from [159] can be seen in Figure 2.2. Note sometimes syllables can be made up of just a single note rather than a group of notes, as seen in Figure 2.2 at about 1 second.

Experiments in chapter 4 evaluate pitch tracking performance on different vocalization types at the syllable level for birds. A description of these syllable types is given here. Syllables tend to fall into one of the following categories:

Whistles: Catchpole et al. [25] describe whistles as the most basic and common type of vocalization. A short whistle of constant pitch appears as a pure, unmodulated frequency trace (see (a) on the spectrogram in Figure 2.3). A sound which drops from a high to low frequency appears as a downward slope (see (b) in Figure 2.3). Whistles can be monotone, upslurred, downslurred, overslurred (where pitch rises then falls) or underslurred (where the opposite is true). Whistles often occur in repetition to form phrases. These phrases can contain a constant series of whistles with each whistle rising or falling in frequency. An accelerating or decelerating series of syllables is also possible [121]. Figure 2.4 shows an example of a decelerating downslurred series of upslurred whistles. Intervals between each whistle will augment over time (decelerating) with each whistle rising in frequency (upslurred whistle) and each whistle syllable will be a lower frequency than the previous syllable (downslurred series).

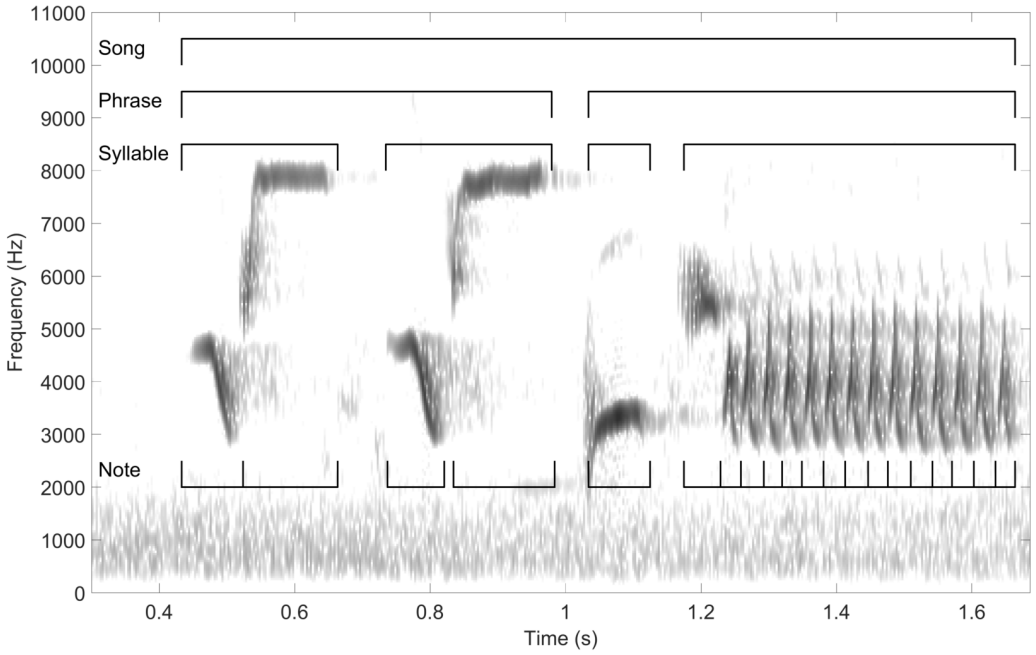


Figure 2.2: A Song Sparrow song showing note, syllable, phrase and song boundaries. This plot is inspired by work in [159].

Hoots: Hoots are just low-pitched whistles, typically less than 1 kHz [121]. These sounds are typical of the voices of doves and large owls.

Trills: Syllables that contain a series of elements or notes which rise and fall in frequency at a rate greater than 10 Hz will be perceived as a trill. Sounds with more rapid modulations are referred to as 'buzzy' sounds. Buzzy sounds are less musical. An example of trilled vibrato and buzzy vibrato can be seen in Figure 2.3 at (c) and (d) respectively. Vibrato range in birds tends to be between 25 - 200 Hz [112].

Noise: Not all bird sounds are tonal or periodic. Noisy sounds are constructed from short bursts of white noise and sound like a click. A noisy example is shown in Figure 2.3(e) and a noisy buzz sound is shown at (f). Noisy bird sounds are likely to be harsh on the ear [121]. As noisy sounds are unvoiced they are excluded from pitch extraction experiments in chapter 4.

Harmonics (or Nasals): Many bird sounds are actually combinations of multiple simultaneous whistles (partials) of different frequencies that the human brain typically perceives as a single sound (because of the mathematical relationship between the frequencies of the different whistles). Harmonic sounds are represented on a spectrogram by a typical ladder pattern and have a noisy spectral quality (i.e. many stacked simultaneous frequencies present). An example

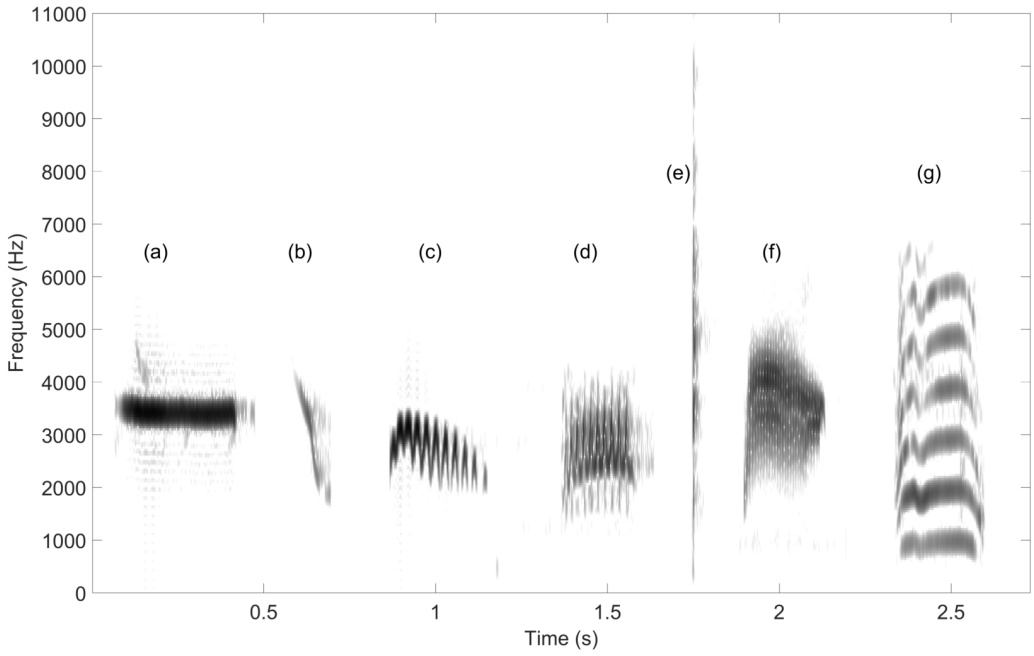


Figure 2.3: Spectrogram of different syllable types based on diagram from [25]. (a) Whistle, (b) Downslurred whistle, (c) Trill, (d) Buzzy sound, (e) Noisy sound, (f) Noisy buzz, (g) Harmonic or Nasal sound.

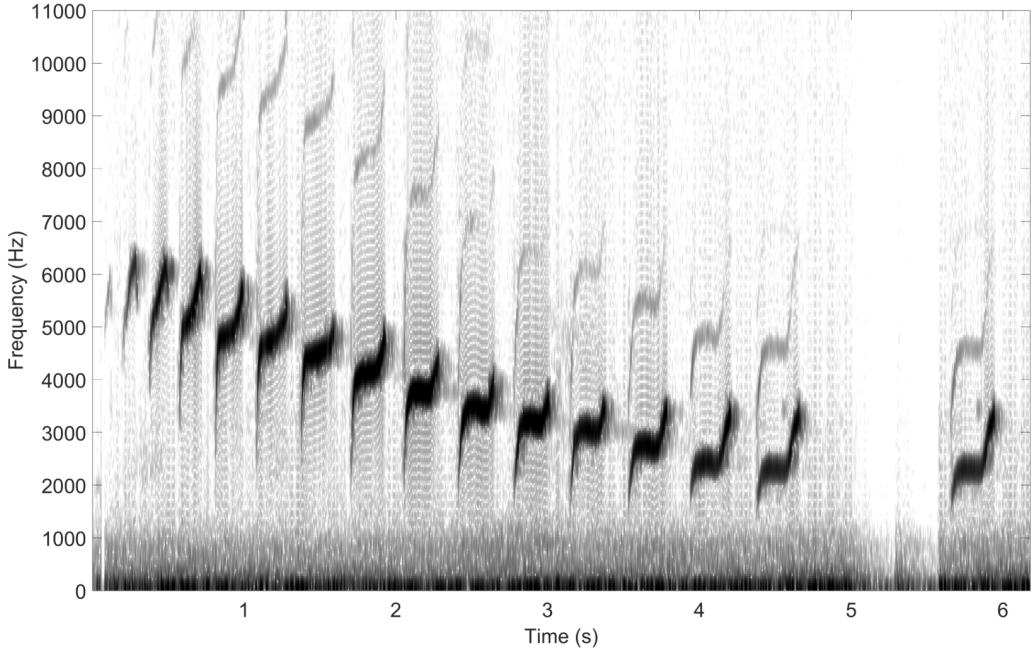


Figure 2.4: Decelerating downslurred series of upslurred whistles. Canyon Wren Song [166].

of a harmonic nasal sound is shown in Figure 2.3(g). The different whistles are called partials because they are only partial components of the sound [121]. If the energy at F_0 or the 2nd partial is prominent, the sound will be soft and melodic because the 2nd harmonic is an octave above F_0. These sounds blend well together. If higher partials have stronger energy, the sounds are more nasally, as the partials tend to clash (the term 'Nasals' is used to refer to these sounds). While whistles can be a pure tone or a combination of strong F_0 with lower amplitude harmonics, nasals refers to sounds with high energy at higher order partials and sound harsher than melodic whistles with harmonics [121].

These nasal sounds are very challenging for pitch tracking, as many nasals sounds have missing harmonics (including missing fundamental) or inharmonic partials [96]. These problems will be discussed in detail later in chapter 4 Section 4.4 and Section 4.5.

Two-voiced sounds: Some birds have the ability to produce sounds with two F_0 values at once using both sides of their syrinx [25]. This results in vocalizations complete with two F_0s, harmonics of both F_0 and heterodyne frequencies [121]. An example of this is shown in Figure 2.5 with a song of a Prothonotary Warbler. Note the labels showing two F_0s (A & B), harmonics (integer multiples of A & B) and heterodyne frequencies (sums and differences of F_0 and harmonics) [121]. Complex two-voiced sounds contrast to many common bird songs that have one main frequency band [149].

While there is scientific literature on the two-voiced phenomenon [177,106,82], its regularity is undocumented. Informally, Zoologists suggest most birds use only one side of their syrinx, some switch between sides during song and few birds use both sides simultaneously. The complexity of pitch tracking and the inaccuracy of ground truth pitch calculations excludes two-voiced sounds from quantitative analysis in chapter 4, but a qualitative evaluation is given in Section 4.5 of that Chapter.

Biphonation sounds: Work in [177] discussed biphonation or two sounds coming from the one source (one side of the syrinx). Biphonation is different to two-voiced sounds. Biphonation refers to a signal produced from one side of the syrinx being modulated to produce more than one simultaneous tone. Biphonation appears as a tone with side-bands. Amplitude modulation can create side-bands, one on each side of the carrier frequency [96]. Some Mockingbirds' vocalizations exhibit a phenomenon known as dual biphonation, where each side of the syrinx produces a tone and each tone is modulated by an unrelated lower frequency (possibly amplitude modulating frequency) resulting in a pair of side bands for each tone giving six lines on the spectrogram three for each side of the syrinx (combination of two-voiced and biphonation phenomena). Biphonation was discussed in great detail in [177] but tracking pitch of such an occurrence is not trialed yet and excluded from experiments in chapter 4.

These syllable types are very broad categories. Some of them grade into one another, and some of them occur in combination, e.g. a note may be simultaneously buzzy, noisy, and

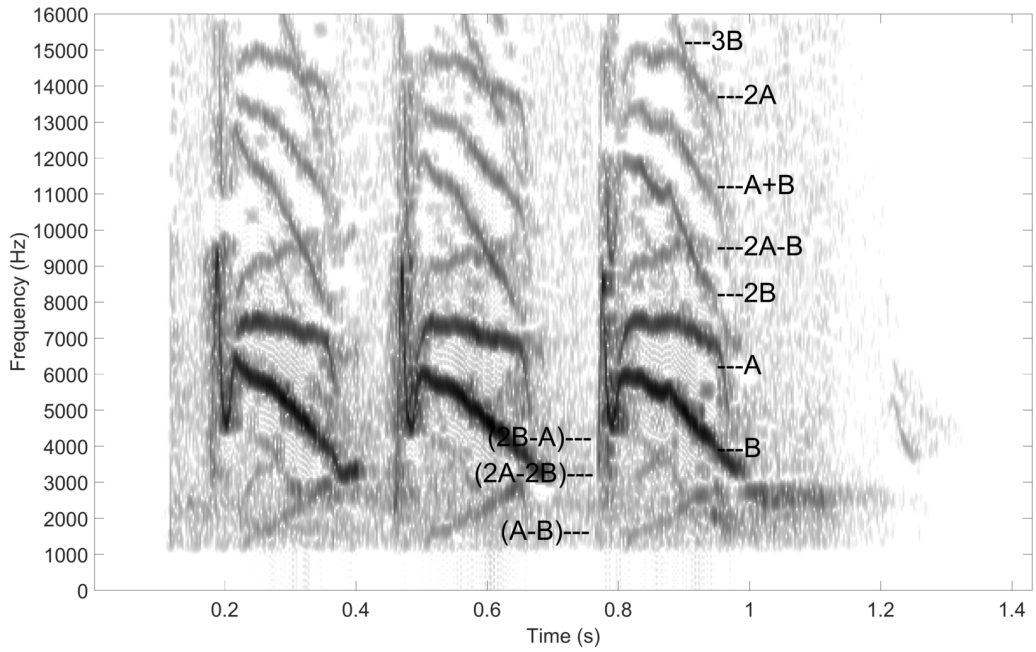


Figure 2.5: Song of single Prothonotary Warbler. Two-voiced bird song with labels showing two F_0s (A & B), harmonics (integer multiples of A & B) and heterodyne frequencies (sums and differences of F_0s and harmonics) [121].

harmonic. Nonetheless this basic vocabulary is very useful when discussing the qualities of bird sounds [121]. These syllable groups are important to work in chapter 4 presented later.

These sections have offered a brief insight into the importance of bird vocalizations. Again a more in depth review of bird song in zoology can be found in [25].

2.4 Automatic acoustic analysis by engineers

Within the speech processing community there has been increasing interest in the analysis of bird song. Signal processing and machine learning techniques can provide much needed assistance to researchers when performing tasks such as species recognition, bird activity detection, phrase classification, song structure analysis and bird song similarity. Speech processing has had a major influence on the automation of bird song analysis. Currently the majority of species classification systems draw on techniques used for speaker identification. Species classification systems involve training models using features (such as Mel-frequency Cepstral Coefficients (MFCCs)) extracted from labeled bird recordings (just like speaker classification). As bird song is produced using a similar source filter mechanism as humans use for speech production, speech processing techniques are quite applicable to bird song analysis. Of course there are differences between bird song and speech, most notably the frequency region of interest is higher for bird

vocalizations. For development in speech processing, clean datasets like TIMIT exist whereas in the bird world research has to start with data recorded in more challenging conditions. Some professional quality bird recordings exist but clean bird data is less common than clean speech data. Recordings generally have to be segmented and labeled post collection. Pattern matching and Voice Activity Detection (VAD) techniques can be used to identify regions of target bird vocalizations in a long field recording. The automation of labeled segmentation saves expert listeners countless hours and removes subjectivity from these tedious tasks. Another issue with bird data is environmental noise. Noise removal methods used in speech processing have also been applied to bird data [31].

Many environmental and scientific applications can be addressed by acoustic analysis of birds. Stowell et al. presented a review regarding acoustic detection of birds in [148]. In this review a number of points were discussed. There is an increasing number of projects now that record bird vocalizations and analyze them. Yet there are still issues to be resolved as a number of projects are yet to be fully automated [97, 21, 45]. Automated data collection means that surveys can take place in locations where it would be too expensive or dangerous to send human observers [97].

As bioacoustics becomes more of a 'big data' research area, terabytes of audio are collected which is far more than can feasibly be inspected manually [148]. Some examples of projects involving bird vocalizations include monitoring of population densities and migration patterns, and monitoring overall ecosystem health. Work in [17], used automatic detection of calling activity to investigate how common seabird colonies are. Sueur et al. in [150] voiced the increased interest in using acoustic indices for biodiversity assessments. Marques et al. in [97] concluded that improvements in automatic detection (and classification) would be desirable, especially with respect to calibration and full automation. Many techniques to date require hand labeled data by expert listeners in the training phase, which is comparable in time and effort with manual survey of a bird population [17]. Parameters need to be set and results need to be post-processed in the majority of cases.

As discussed in this section there are a number of applications for which automatic bird song analysis is desirable. Two major applications will now be reviewed, bird species classification and bird event detection.

2.5 Bird species classification

Automatic classification of bird species using their sounds is a rapidly expanding field. Work to date has focused on species classification, inspired by automatic speaker recognition, using recordings of calls and song [59, 165, 58, 56, 57, 51, 123, 145, 55]. Trifa et al. [165] presented an HMM based approach to distinguish song from five species of antbirds that share the same territory in a rainforest environment in Mexico. When only clean recordings were used species recognition was 99.5% and when noisy data was included this rate dropped to 90%. Besides

quality of recordings, performance was found to be heavily influenced by a number of factors, such as size of training set, the feature extraction method used and the number of states in the HMM. The results in [165] show HMMs are an effective method for discriminating bird species. MFCCs and Linear Prediction Coding (LPC) were used with the use of MFCCs outperforming LPC.

Fagerlund et al. [50] used MFCCs and a decision tree with Support Vector Machine (SVM) classifier at each node to perform classification between species (6 species for one dataset and 8 species for a second dataset). An iterative procedure which estimates the background noise level as it converges was used in pre-processing. Data was manually segmented into syllables and separated into test and training subsets. The experiments in [50] showed that using an SVM classifier works well with MFCCs. This evidence supports the general consensus that the MFCCs are high performance features suitable for bird song signal processing.

Lee et al. in [88] presented a bird classification approach using a feature set derived from static and dynamic two-dimensional Mel-frequency cepstral coefficients. Best classification accuracy was 84.06% for classification of 28 bird species. Song Scope was developed by Wildlife Acoustics led by Agranat [1, 173]. This software can perform bird song recognition from their vocalizations using HMMs and feature vectors similar to MFCCs with significantly less human effort required than manual inspection of recordings.

Shaw [145] in 2011, presented a comparison of different bird song classifiers as part of his undergraduate dissertation. Classifier performance of Song Scope, SVM, GMM, neural networks, HMM and a tandem neural network/HMM system using 15 species of bird were compared. Neural network, GMM and HMM gave better results while SVM and Song Scope performed poorly. While this is not an exhaustive comparison, findings strongly motivate the use of GMMs and HMMs for species classification of birds.

Gracierena et al. [57] explored front-end feature extraction parameters to improve bird species recognition. The baseline experiments used GMMs with MFCCs. They found improvement when increasing the spectral bandwidth and increasing the number of filter banks. No improvement was found when switching the filter bank distribution from the perceptually based Mel-frequency scale to a linear frequency scale. In addition, no improvement was found when the time resolution was either reduced or increased. On the other hand, the best time resolution was found to be species dependent. Significant improvements were found from a species-specific combination of different front-ends with different time resolutions. Most improvements were minimal by varying parameters. Commonly used parameters for MFCC extraction of speech data will work quite well for bird data, e.g. Mel-spaced, 25 filters, 25 ms window, 10 ms frame rate, bandwidth of 0-0.5*Sampling Rate ($\boldsymbol{F_s}$).

Birds with tonal song are comparable to time-varying sinusoids. Jančovič et al. developed a sinusoid tracking system for speech in [65], then applied it to bird vocalizations in [64] and again with further refinements to improve its accuracy on bird segmentation and classification in [67]. The annotation came in the form of frequency track identification by sinusoids using HMM

models for noise and sinusoids. These sinusoid features were then classified as 1 of 30 species using HMMs. Since conventional MFCCs capture the entire frequency band, they are prone to background noise and presence of non-target vocalizations in other bands. The frequency track feature approach, if extracted well, is preferable when processing field recordings of bird vocalizations which contain noise and non-target sources. Experiments in [64] report these tonal features outperform MFCCs on noisy and mismatched data using GMMs, with comparable performance between them on matched data. Experiments in [67] report that bird species recognition accuracy improves from 68% to 84.3% when using HMMs instead of GMMs with frequency tracks. MFCCs have the advantage that they can be applied to aperiodic and periodic signals [22], but MFCCs are not noise robust in mismatched conditions [71]. The additional modeling of temporal information when using HMMs instead of GMMs is useful especially when using song, which is more temporally complex than calls which are simpler. Sometimes the lack of data precludes the use of HMMs and GMMs are used instead.

Cai et al. [22] investigated neural network performance of bird species recognition in 2007 with different pre-processing methods and sets of features. 14 species from the Australian subtropical east region were used in the classification experiment. The main features used were MFCCs and linear scale cepstral coefficients. The main classifier was a multilayer perception neural network. Neural networks with 160 hidden units using MFCCs gave an accuracy of 86.8% while linear cepstral coefficients gave an accuracy of 84.6%. This was another example of MFCCs being successfully utilized to identify bird species. The exact data size was not given in [22] but it is presumed large as this is a requirement of neural networks.

Large and structured communities of nature observers (e.g. xeno-canto.org [166]) have started to produce outstanding collections of multimedia records. Unfortunately, the performance of the state-of-the-art methods on such data is still far from reaching real world requirements. This prompted LifeCLEF to introduce annual challenges for researchers to improve technologies for applications in multimedia information retrieval and classification in the living world [69]. One of these challenges was BirdCLEF an audio based bird identification task. This challenge significantly increased the number of species for classification using real world data collected by thousands of recordists, moving to a more usage driven and system oriented benchmark [69]. Overall, the task is more difficult than previous benchmarks because of the higher confusion risk between the classes, the higher background noise and the higher diversity in the acquisition conditions (devices, recordists used, contexts diversity, etc.). It therefore produces substantially lower scores with progression towards building real-world generalist identification tools [69]. Tóth and Czeba described their entrant to BirdCLEF 2016 in [162]. They used a Convolutional Neural Network (CNN) based deep learning approach to the bird song classification challenge. Spectrograms were pre-processed to remove regions suspected to be non-target information, and used as input to the neural network system. Training and testing data contained about 24,000 and 8,500 recordings respectively, belonging to 999 bird species. Using recordings in which there was a dominant singing bird in the foreground with some other ones

in the background, the mean average precision was 33.5%. For recordings with only one bird singing, mean average precision increased to 42.4% (labeled as BME TMIT Run 4 in [147]). According to official results the system in [162] came 4th in the BirdCLEF 2016 competition of 6 entrants.

Sprengel et al. [147] were the winners of BirdCLEF 2016. They also used a CNN approach. Their network contained 5 convolutional layers, each followed by a max-pooling layer. The dense layer contained 1024 units and the soft-max layer contained 1000 units, generating a probability for each class. Batch normalization was used before every convolutional layer and before the dense layer. The convolutional layers used a rectify activation function. Drop-out was used on the input layer (probability 0.2), on the dense layer (probability 0.4) and on the soft-max layer (probability 0.4). As a cost function the single label categorical cross entropy function (in the log domain) was used. Their methodology included data augmentation using time and pitch shifting to increase data size. Their approach surpassed state of the art performance when targeting the dominant foreground species. Using recordings in which there was a dominant singing bird in the foreground with some other ones in the background, the mean average precision was 55.5%. For recordings with only one bird singing, mean average precision increased to 68.6% (labeled as Cube Run 4 in [147]). When background species were taken into account, other approaches performed almost as well as [147]. When no foreground species were present one other approach was able to outperform [147]. This was not surprising considering their data augmentation approach and pre-processing method which enhances foreground signals.

Another related topic in bird song analysis is syllable or phrase classification [72,155,154,79] and song structure [138]. Kaewtip et al. [71] presented a phrase classification system that can be utilized with limited training data and data which is corrupted by background noise. The algorithm utilizes Dynamic Time Warping (DTW) and prominent (high-energy) time-frequency regions of training spectrograms to derive templates. The performance of this algorithm was compared with the traditional DTW and HMM methods under several training and test conditions. DTW works well when the data is limited, while HMMs do better when more data is available, yet they both suffer when the background noise is severe. The proposed algorithm outperforms DTW and HMMs in most training and testing conditions, usually with a high margin when the background noise level is high. For this case of phrase classification, classes had very few examples. When there are only 5-10 examples of a phrase, a pattern based DTW approach is feasible but there is simply not enough examples to accurately train a HMM or GMM. Tan et al. [154] evaluated bird phrase classification for the same species used in [71] using a sparse representation classifier. The system outperformed nearest subspace and SVMs.

A growing issue in ornithology is to determine how similar calls or songs between subspecies of a single bird species are. This has important implications for the correct taxonomic classification of populations. A theory that bird vocalization difference is comparable to how separable bird species are by their vocalizations was investigated by Harte et al. in [59]. Work in [59] was the first to examine the issue of call similarity in this context. That paper demonstrated that

classifier performance is related to similarity but is not a quantifiable indicator. Alternative methods of measuring bird vocalization difference are given in Section 2.8. Experiments in [59] used GMMs trained with MFCCs. Of course bird vocalization difference will depend on the metric used.

The study of species identification and classification is not exclusive to birds. Elephants [33], dolphins [80], marmots [14], whales [139], grasshoppers [29] and monkeys [77] to name a few have been subjects of acoustic analysis. Generally much of the work that is carried out does not build on previous work, which leads to difficulties when comparing research [145].

This brief background shows that MFCCs and GMMs/HMMs are quite useful for the application of bird species classification. Their use in other bird song analysis applications is also presented. Work in this thesis uses MFCCs and GMMs as baseline for bird classification in chapter 3. They are also used for detecting bird events within field recordings of bird vocalizations in chapter 6.

2.6 Detection of bird events

Work in chapter 6 introduces a novel speaker diarization approach to segmentation of long bird recordings sometimes referred to as detection of bird events. To apply machine learning procedures to bird vocalizations (such as species classification), training and test data needs to be segmented into phrases, calls or songs. Previously when working with continuous recordings for species classification, signals had to be segmented by manual inspection of spectrograms [19]. This can be laborious and prone to subjectivity. Work in [59] used data which was segmented into one call per file manually. The data in [71], used for phrase templates, used annotation files containing phrase information labeled by expert listeners. The test data for the HMMs also had to be pre-segmented in [71]. The Mexican antbirds data in [165] were divided into one song per file. All of these approaches required some sort of manual pre-processing of data. Species classification normally involves pre-segmenting longer recordings into individual test examples. Each test example is then compared to pre-trained models and a prediction is made. Many bird classifier systems employ MFCCs with GMMs or HMMs. The species classification framework could be modified to segment long bird recordings. For segmentation, models are trained for birds and background sounds. Each frame will be compared to the models and a predicted output will classify events into bird song or background. Although the structure remains the same for this application, the task has changed from identifying bird species (species classification) to detecting bird vocalizations and non-bird events (comparable to acoustic event detection).

Some automatic techniques have been employed, e.g. using an energy based threshold to segment the data in time (and sometimes in frequency). An energy based threshold was used in [18, 50, 146]. This approach can be problematic when the amplitude of target and non-target individuals are similar within the same recording. Work in chapter 6 presents approaches which overcome the shortcomings of energy based segmentation.

Many biological monitoring projects also rely on acoustic detection of birds. Despite large datasets, this detection is often manual or semi-automatic, requiring manual tuning/post-processing [148]. This manual pre-processing is time consuming, tedious and subject to the opinion of the expert listeners. An objective of work in this thesis is to help make bird song analysis fully automated. This involves automatic segmentation of bird recordings using bird event detection which presents the task of identifying song presence with onset and offset time.

These diverse uses have broadly similar requirements, but can differ in the exact precision of detail required. Applications may require exact start/end times, vocalization counts, species counts, individual counts or simply presence/absence of phrase or species [148]. The identification of when a bird is vocalizing during a long field recordings is referred to as 'Detection of bird events' here. With accurate segmentation of longer bird recordings into a neat one song per file format, this data can then be passed to a classifier to identify species or phrase class. This latter stage is referred to as 'Bird song classification'. For recordings that contain only one bird with little background noise and no non-target vocal sources the detection of bird events is not necessary, but for noisy real world recordings corrupted by environmental noise, non-target birds and other animals, it is necessary to extract the bird acoustic event information. Without this segmentation, periods of silence or periods containing other forms of acoustic information will be incorrectly included when training a class model or when extracting temporal or spectral features for bird song analyses.

Stowell et al. in [148] discussed a number of different task paradigms along with their characteristics. Their differing characteristics have strong implications, both for which computational approaches are appropriate and for the practical feasibility of annotation and evaluation.

2.6.1 Presence/absence

The most basic form of bird detection is the simple estimation of presences/absence in a given sound clip, where a detector outputs zero if none of the target species are detected and a one otherwise (see Figure 2.6a) [148]. This is useful when the task is to identify if a bird is present in a recording or not. However it has low temporal detail and multiple events are merged. It could be used as a first step in manual segmentation, where files that contain no bird vocalizations are removed before manual inspection. For applications like filtering a large dataset, annotated fine detail is often unnecessary since the purpose is simply to help a person or machine remove non-target files and focus on target files.

2.6.2 Voice Activity Detection (VAD)

A more useful style of bird detection is analogous to Voice Activity Detection (VAD) in speech processing [127, 53, 176]. Time is partitioned into positive and negative regions by short time energy levels using a threshold or model based approach. In VAD applications for speech, there is typically a dominant target speaker of interest, while in field recordings of birds there

can be multiple individuals of the target species. Thus a positive region may contain one of many vocalizations [148]. For bird song analysis VAD is usually preceded by some kind of noise reduction and often applied to bandlimited regions of interest [48, 100]. VAD methods of segmentation retain the simplicity of the presence/absence method but adds valuable timing information of events as shown in Figure 2.6b. Recordings where two birds of the same species are vocalizing is a problem similar to separating two target speakers speaking at the same time. Work here only used data that contained a single target per recording.

Energy based detection can be performed by a bioacoustics software called Raven [26] and has been used in various surveys [148]. Although energy based time domain segmentation has been successfully applied to low-noise, single-bird recordings, audio from field recorders often contains too much noise for such methods, so a more robust segmentation method is usually required [109].

A more complex method of VAD which results in a similar style output of song start and end times uses HMMs. HMMs have been widely used for sound sequence analysis in speech and have particular appeal of temporal flexibility that goes beyond temporal matching [148]. Duan et al. in [48] used Song Scope software [1] which utilizes HMMs for bird detection in a comparison against threshold methods used by Raven. Song Scope is aimed at detecting call structures, which is a different approach to Raven. The Song Scope classification algorithms are based on HMMs using spectral feature vectors similar to MFCCs as these methods have been proven to work effectively in robust speech recognition applications [2, 48]. However, this approach is very sensitive to the purity of syllables. If syllables are polluted by non-target species or background noise, the model is very sensitive, thereby affecting the recognition accuracy. Using Song Scope effectively, requires some background knowledge of signal processing to understand and setup the parameters. Song Scope also supports batch processing to deal with large datasets [48]. Graciarena et al. [57] used a simple VAD system, with acoustic models trained with bird vocalization data to segment their bird recordings. Bird vocalization segments were extracted from CD waveforms. Segments of bird vocalizations were identified, with background noise and very short call segments discarded. Wei Chu [31] also used HMMs to detect Robin syllable boundaries.

Most of the recordings from the literature contain one target bird species per file corrupted by non-target sounds. The most important function of this style of bird detection is to identify the main target vocalizations and discard the rest of long field recordings. Work in chapter 6 presents a speaker diarization approach to this VAD application for birds. Each recording is assumed to have one target bird. Song from that bird is then identified and temporal boundaries assigned. Background sounds and non-target song are rejected. This system does not depend on species of the target.

Work in chapter 6 concetrates on the VAD style of segmentation but other styles of bird detection are discussed here for completeness.

2.6.3 Polyphonic Segmentation

Benetos et al. in [10] proposed a system to deal with overlapping acoustic event detection, which models the temporal evolution of sound events. The temporal succession of the templates was controlled through event class-wise HMMs. As input time/frequency representation, the Equivalent Rectangular Bandwidth (ERB) spectrogram was used. Experiments were carried out on datasets of office sounds containing multiple sound sources at once generated using an acoustic scene simulator, as well as real and synthesized monophonic datasets for comparative purposes. This could be used to segment bird recordings not only in time but in frequency as well if different birds were vocalizing in different regions of the spectrogram at once. A graphical representation of this detection style is given in Figure 2.6c. Other methods such as Non-negative Matrix Factorization (NMF) [40] are designed around having multiple source types each of which can be active/inactive, yet within each source type they do not segregate signals that may overlap or come from multiple individuals. NMF is an algorithm for describing data as the product of a set of bases and a set of activations, both of which are non-negative [40]. It is useful for finding parts-based decompositions of data. Since all components are non-negative, each basis contributes only additively to the whole. This promotes a solution in which high energy foreground events and constant low-level background energy may be described by different bases, and therefore separated in the feature representation [40]. This type of approach can deal with multiple birds if their vocalizations stay in different frequency bands. These approaches have not received much attention in research on animal sound detection [148].

2.6.4 Time-frequency Boxes

When bird sounds are compact in time and frequency, events can be described via time-frequency boxes. These can be annotated intuitively by drawing boxes on a spectrogram and detected using template-matching methods (see Figure 2.6d). This method is not common in speech and music analysis because in those cases the signals of interest are often broadband, consisting of harmonic stacks and noise [148]. Bird sounds for the most part are bandlimited and for which the time-frequency box segmentation method should work quite well, but for sounds with significant energy across a range of harmonics, there may be a tendency to exclude higher harmonics or to create large regions containing many subbands with no energy from signal of interest. This may impact accuracy for some species [148].

Spectrogram cross-correlation is a method which uses one or more example sounds as templates to scan a spectrogram for regions with strongly-matching profiles by cross-correlation. These templates tend to be species specific. Borker et al. in [17] used XBAT [27] to perform cross-correlation to detect calls of a seabird colony with positive results.

2.6.5 Time-frequency Blobs

A more complex form of segmentation stems from object-detection in image processing. This method relies on detecting arbitrarily shaped regions on a spectrogram shown in Figure 2.6e. Neal et al. [109] proposed a supervised time-frequency audio segmentation method using a Random Forest classifier to extract syllables of bird calls from noisy signals. Using the spectrogram units as features the random forest method outperformed energy based thresholding. Often each detected event is required to be a single fully-connected region ("blob detection"), which is problematic for harmonic sounds since harmonics may then be detected as separate events. Manually labeling data at this resolution is labor-intensive, and it is rare that the final downstream applications require such detail [148].

2.6.6 Time-frequency Sinusoids

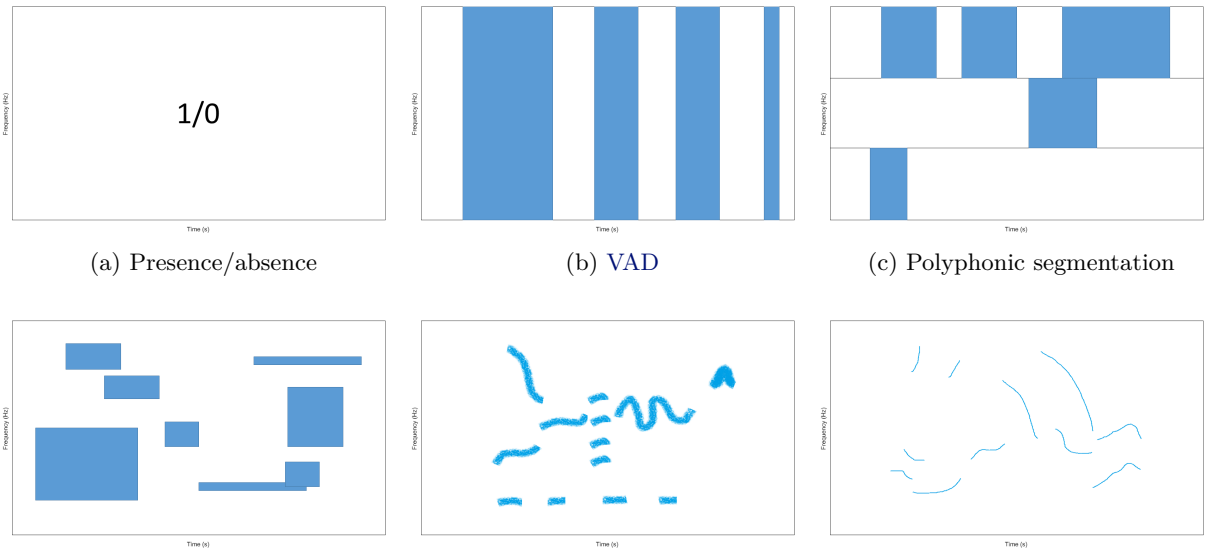
As briefly mentioned earlier, tonal bird sounds are comparable to time-varying sinusoids. Jančovič et al. developed a sinusoid tracking system for speech in [65], then applied it to birds in [64] and again with further refinements to improve its accuracy on bird segmentation and classification in [67]. The annotation came in the form of frequency track identification by sinusoids using HMM models for noise and sinusoids (see Figure 2.6f). Again, harmonics might be detected as separate events, however some models are able to conjoin harmonics into unified tracks. Also note that a sizable number of bird sounds are non-tonal, and the very rapid pitch modulation of some bird sounds can cause problems for standard frequency trackers [148]. Although, the system in [64] is not affected by rapid pitch modulations, for instance, a frequency changed from 8950 Hz to 5850 Hz during approximately 20 ms of a song example is given in [64]. Despite these fast frequency variations, the sine-distance showed good detection, well able to track the bird singing frequency [64]. The sinusoidal detection method indicates which spectro-temporal points were detected as sinusoids. This can be considered as an initial segmentation of the acoustic scene.

The task of bird activity detection is important to many aspects of bird song analysis. A number of bird detection approaches have been discussed. They provide varying levels of detail for downstream tasks, but they also enable different sets of technical solutions, and imply different amounts of manual labor to annotate and evaluate [148].

2.6.7 Diarization applied to bioacoustics

Kojima et al. in [81] presented a semi-automatic annotation scheme for bird song analysis. Robot audition[3] methods including sound source separation, localization, separation and identification were typically studied separately and utilized in a cascaded manner with poor performance. Work in [81] suggested a more integrated setup of robot audition within a Bayesian framework.

[3]The ability of robots to listen to several things at once with their own "ears", that is, robot audition, is an important factor in improving interaction and symbiosis between humans and robots [113].



(a) Presence/absence (b) VAD (c) Polyphonic segmentation

(d) Time-frequency boxes (e) Time-frequency blobs (f) Sinusoid tracks

Figure 2.6: Graphical representation of bird event detection methods. These are based on figures from [148].

This method doesn't require pre-trained models removing the need for a large set of annotated data. A percentage of a recording was annotated by specialists and the rest was automatically annotated using their system. Their data contained up to 8 birds vocalizing in a given recording (recorded with a microphone array). When 0.6 of the recording was annotated the remaining data was automatically annotated with 67% accuracy, which is good considering the difficulty of their task. This method could possible be adopted to bird segmentation with less non-target competition.

Ciira wa Maina [168] was the first to apply speaker diarization techniques to bird song analysis. In [168], speaker diarization was used to identify change points and cluster segments into classes. One class was assumed to be non-bird sounds, while the rest of the clusters were taken to indicate the number of species present in the recording. Thus the system explored the relationship between clusters and species counts in recording of 1-6 birds. The correct number of species was obtained for 76 of 179 recordings. Work in chapter 6 utilizes a similar speaker diarization setup but instead now applied to bird song segmentation. The ability to identify change points throughout a recording is central to the potential strength of such an approach. Experiments evaluate segmentation accuracy of target song (with time boundaries) from field recordings with a single target bird corrupted by other audio events. Trained models can refine boundaries and remove non-bird events detected as song. The system is compared to baseline segmentation using HMMs.

Bird Taxonomy	Taxa Counts of Woodland Park, Washington, USA	Example of Mexican Antbird Taxonomy
Kingdom	Animalia	
Class	Aves	E.g. Bird
Order	45	E.g. Antbird
Family	≈235	E.g. Thamnophilidae
Genus	≈2340	E.g. Thamnophilini
Species	≈10000	E.g. Dusky Antbird
Subspecies	≈30000	E.g. Dusky Antbird of X

Figure 2.7: Bird taxonomy example taken from a presentation by Woodland park Washington, USA [174]. A specific taxa example for the Dusky Antbird (DAB) species is also given.

2.7 Bird Taxonomy

Species are the fundamental units of biology, ecology and conservation but progress in these fields is hampered by widespread taxonomic bias and uncertainty [160]. Taxon (plural 'Taxa') refers to any group of organisms. Taxonomic ranks from high to low are, Life, Kingdom, Class, Family, Genus, Species [4] and Subspecies [5] respectively (with some minor ranks excluded). Figure 2.7 shows information from a presentation given by a zoo in Washington, USA which describes bird taxonomy. Examples for different taxonomic levels of the Dusky Antbird (DAB) species are given in the right hand column. A subspecies of DAB would be a subset population of DAB species found at a location X that slightly differs from the species as a whole.

Decisions on avian taxonomy were once uncontested matters but over the last few decades species lists have been reviewed [160]. While biodiversity can be divided into a range of categories from genes to ecosystems, it is the species category that underpins much of biology, ecology and conservation [99, 93, 43, 160]. Species are crucial to conservationists and policy-makers, who use them as units for prioritizing action and formulating law, and who therefore require species delimitation to be consistent and transparent [134, 35, 124, 54, 160].

Reclassification of any bird is a task which requires many characteristics to be examined such as morphological and plumage differences. The distinctiveness of their song is equally as important, hence bird song is crucial to biodiversity studies. Automatic acoustic analysis

[4] A species is commonly defined as the largest group within which interbreeding produces viable offspring.

[5] A subspecies is a subgroup below the level of a species. It is a group which can interbreed successfully with other subspecies, but does not do so in practice (e.g. due to geographical isolation). Species conservation precedes subspecies conservation which is why correct taxonomy is so important.

presented in this thesis contributes to the reclassification process.

2.8 Quantifying difference of bird vocalizations

As mentioned in Section 2.2, bird vocalizations are a strong discriminator when identifying species. Ornithologists are interested in determining how similar, or different, voclaizations from subspecies of a single bird species are. Correct taxonomic classification of bird populations is critical to conservationists. Groups of birds considered as the same species, but as different subspecies, will typically have more similar vocalizations than groups of birds considered unrelated species. The task of classifying subspecies is more complex than species classification as classes are more confusable. Whilst ornithologists will use a combination of genetic and morphological evidence (e.g plumage patterns) to help them in this task [59, 57], the usefulness of vocalizations is also now strongly accepted. Systematic ways to formulate difference between groups of birds using their vocalizations is thus desirable for ornithologists. Work in chapter 3 presents a system which measures bird vocalization difference using pitch contours. This system is inspired by dialect distance work in speech.

Engineers have used bird vocalizations to classify bird species automatically. In zoology bird calls and songs are used to gain vital information about a population. Vocalizations together with plumage patterns and other forms of morphological information tell a lot about bird species [59, 57, 145, 89, 88, 129, 137]. Accurate vocal analysis can be used as evidence in cases where zoologists wish to reclassify a population as a new species. Work in this thesis helps automate this vocal analysis.

Standard approaches in ornithology to bird song analysis are laborious, subjective and can lack repeatability. McKay et al. in [101] examined song in making a case for the Bahan subspecies of the Yellow-throated Warbler to be reclassified as a distinct species. Song divergence was important evidence in the reclassification process. Comparisons were on the basis of visual inspection of spectrograms. Sangster et al. in [137], described a new species of owl, known as the Rinjani Scops Owl, based on analysis of vocalizations. In both [101] and [137] various features were measured, like amplitudes at certain frequencies, number of syllables and phrases, pitch slope and frequency. One of the goals of this thesis is to quantify a measure of difference between bird populations using their vocalizations, that is accurate and repeatable to aid ornithologists in their study of bird biodiversity.

While the importance of vocalizations in mate choice and species recognition is well documented (specifically by Catchpole et al. in [25]), quantitative systems to evaluate difference are few and far between. The task of bird population difference analysis using vocalizations is relatively unchartered with only a few works to date [59, 116, 105, 137]. It is now widely accepted in the ornithology domain that acoustic signals give early clues of species distinction [86]. If two populations with a common origin are isolated, one can expect that the songs of each will accumulate modifications independently. Being able to detect those changes can help infer

population histories and relationships [128].

Prosodic features, like pitch, have been used to quantify difference in bird populations. Quantitative measures of acoustic similarity were used to investigate patterns of shared vocal behavior in social species by Meliza et al. [105]. Pitch or Fundamental Frequency ($\boldsymbol{F_0}$) based methods performed best at separating distinct categories of superb starling calls. Understanding how vocalizations are shared among individuals of the same species requires quantitative methods for measuring how acoustic features vary across groups and individuals [105]. $\boldsymbol{F_0}$ analysis is also commonly used by zoologists. Tobias et al. [160] developed a system of standardized criteria for species delimitation in birds. Acoustic evidence of song structure like maximum frequency, minimum frequency, bandwidth and peak frequency were used. This paper will be discussed in great detail later as work in chapter 5 automates the acoustic feature extraction process in [160]. Lachlan et al. [85] included pitch to evaluate chaffinch song. These examples highlight the growing interest in accurate acoustic analysis of bird song. For tonal and harmonic vocalizations, $\boldsymbol{F_0}$ is a particular useful basis for comparison as at least some birds perceive tonal sounds in the same way humans do [105]. Thus there is growing agreement among ornithologists that pitch analysis of bird vocalizations is useful, as many avian calls and songs are tonal [157, 105].

To better understand bird song analysis, engineers must take a glimpse into how zoologists study birds. Before attempting to quantify bird population difference, engineers must also understand how and why zoologists and ornithologists analyze acoustic signals. One of the largest studies of its kind was reported by Tobias et al. in [160]. Tobias et al. gave a detailed account of bird population differences in [160]. In [160], a system of standardized criteria for species delimitation in birds was constructed. Mean levels of difference in multiple phenotypic (organism's observable characteristics) characters including biometrics, plumage and voice for 58 sympatric (occurring within the same or overlapping geographical areas) or parapatric (populations with divergence between different ends of the same geographical regions) species-pairs from 29 avian families were calculated. While recognized that species limits are in many cases inherently arbitrary, Tobias et al. argued that this system can be applied to the global avifauna to deliver taxonomic decisions with a high level of objectivity, consistency and transparency [160].

The main goal in [160] was to produce a system that is simple and rapid enough to be applied to all birds, quantitative enough to satisfy the need for rigor and repeatability, and transparent enough to allow taxonomic decisions to be traced and evaluated. The approach described in [160] is based on measuring phenotypic divergence in undisputed species to establish thresholds for a taxonomic scoring system. An organism's physical appearance and behavior, influenced by genetics and environmental factors is referred to as a phenotype while genotype is the genetic makeup of an orgasm. The term 'trait' describes a broad class of phenotypic difference like plumage, song or biometric, and 'character' is used to describe any diagnostic difference identified by a cross-taxon comparison. It is assumed that characters can be described, measured or counted and are consistently present in members of the same age and/or sex class of a population, on the basis of which that population can be classified as a taxon. The most

relevant characters for work in [160] are morphological (e.g. size and shape), visual (e.g. plumage color and pattern), acoustic (e.g. pitch and pace) and behavioral (e.g. courtship display or nest type). It is worth observing that allopatric disjunction, however great, is not in itself a character, and should not be taken as evidence of species status [160].

There are many advantages to using phenotype to delimit species. In particular, phenotypic characters are often determined by multiple genes, and thus an assessment of numerous phenotypic characters are more likely to reflect divergence across the genome than are molecular methods dependent on limited sampling of locations [160]. Early taxonomic work emphasized the importance of divergence in morphological and plumage traits, while ignoring other forms of phenotypic divergence, such as vocal and behavioral traits. The work in [160] presents a system with a more complete set of phenotypic characters in a quantitative framework. The inclusion of acoustic features is what makes this work appealing to signal processing engineers, hence why work by Tobias et al. is so important to this thesis.

Tobias et al. discussed not only acoustic traits but the use of behavioral, morphological and plumage traits to set species limits in [160]. Previous divergence assessments, like British Ornithologists' Union (BOU) guidelines presented in [60], purely tried to identify lineages that have separated. The goal of work in [160] was to identify those that have separated in such a way that they are likely to be reproductively isolated. From this perspective, the most appropriate taxonomic characters will be those contributing directly to reproductive isolation. A single character that makes such a contribution must inevitably have greater significance than several characters that do not. In other words, minor differences in gene sequences, or in characters like bill length and shade of plumage coloration, explicable as adaptations to local environments, are perhaps unlikely to represent barriers to interbreeding, whereas the opposite will be true of differences in characters related to courtship or species recognition [160]. By the same logic, large differences in any phenotypic trait are more likely to reflect lineage separation and reproductive isolation than small differences. Previous criteria generally gloss over vocal or behavioral differences, and lump together minor and major phenotypic differences as characters with equal weighting. The system in [160] focuses more attention on mating signals such as songs and displays, and weights all characters according to the degree of their divergence.

Taking all that into account, how does one decide what taxa are to be measured and what populations or subspecies are to be compared? When species have only two distinct subspecies the choice is clear-cut. In the case of species with multiple subspecies the pattern of character distribution can become highly complex [160]. Subspecies can vary from very similar to very different. Multiple comparisons between different populations are required. Tobias et al. suggests as a general rule that phenotypically close taxa should be compared even when geographically distant. Tobias et al. also suggests that, at least for conservation purposes, individual taxa be split off if they meet criteria, rather that waiting for a full review of subspecies.

Biodiversity tends to segregate neatly into discrete species at any given locality, and therefore taxonomic confusion is negligible in sympatry (in biology, two species or populations are

considered sympatric when they exist in the same geographic area and thus regularly encounter one another). However, the wider picture is greatly complicated by geographical variation and the gradual evolution of reproductive isolation in allopatry (allopatric populations of related organisms are unable to interbreed because of geographic separation). Lineages diverge slowly, and taxonomy involves placing cut-offs somewhere along the transition from populations to species [160]. Many primitive speciation decisions were subjectively made and may need to be reviewed.

In summary bird species limits are dependent on a number of different traits. Acoustic analysis by engineers hopes to aid ornithologists in their analysis of species limits by measuring important spectral features (like pitch) and temporal features (like song boundaries) accurately and efficiently. While this introduction to phenotypic divergence of bird traits may be excessive for the task of acoustic analysis, it is beneficial to know where acoustic evidence fits in to the assessment of species limits. Work in chapter 3 and chapter 5 present automatic methods of quantifying difference in bird vocalizations to aid assessment of species limits.

2.9 System Implementation

This section reviews some of the commonly used techniques in speech processing and bird classification.

2.9.1 Mel-frequency Cepstral Coefficients (MFCCs)

The features commonly used to train GMMs in speaker recognition are Mel-frequency Cepstral Coefficients (MFCCs) [41]. MFCCs have been successfully used as features for many speech processing applications, including speech, speaker, emotion and language recognition [75]. Recently, almost all speech recognition systems use MFCCs (or Perceptual Linear Prediction (PLP)) as features. As bird vocalization production can be modeled as a source-filter mechanism (as can speech production), MFCCs can also be used to represent bird vocalizations. There are a number of advantages to using MFCCs. The first is that the Mel scale frequency is consistent with human auditory perception. Mel-frequency features have been proven to be able to serve speech recognition systems better than linear frequency features [22]. Secondly MFCCs can be extracted from aperiodic and periodic signals [22]. And a third advantage is cepstral coefficients can achieve significant data reduction with reasonable information loss [22].

MFCCs have been also used in bird classification in [56,57,88,165]. The process of MFCC extraction starts with an acoustic signal being windowed (usually Hamming). The windowed signal is transformed into the frequency domain using the Fast Fourier Transform (FFT). The resulting magnitude spectrum, the *spectral envelope* of the speech (or bird song) signal, is a descriptor of the resonance properties of the vocal tract, which carry bird species-specific information [75]. This spectrum is filtered by a set of filters, with bandwidths and spacing determined by the

perceptually-motivated *Mel* scale. Lower frequencies are given a stronger weighting. The log of the output of the Mel-filter bank is then passed through a Discrete Cosine Transform (DCT), which decorrelates the features (this allows diagonal covariance GMMs to be used). Finally an MFCC feature vector is obtained by retaining the first 10-20 coefficients. Coefficient c_0 represents the energy of the frame and is often not included in the feature vector [75].

2.9.2 Classification using Gaussian Mixture Models (GMMs)

Gaussian Mixture Models (GMMs) have been the dominant modeling approach in speaker recognition for some time [75]. Finnian Kelly, a PhD graduate from the Sigmedia laboratory at Trinity College Dublin (TCD), gave a detailed description of GMM speaker recognition systems in [75]. A brief summary of that description is provided here.

Audio recordings are used to train a GMM for each class (speaker for speaker recognition or bird population for bird species classification). A GMM [131] is a weighted sum of M multivariate Gaussian components (mixtures), allowing it to model an arbitrary distribution of observations. The Likelihood of an observation x given a GMM denoted by λ is:

$$p(x|\lambda) = \sum_{m=1}^{M} w_m p_m(x) \tag{2.1}$$

where x is a D-dimensional vector, w_m is the weight of the m^{th} Gaussian component $p_m(x)$,

$$p_m(x) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_m|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(x-\mu_m)'\Sigma_m^{-1}(x-\mu_m)\right\} \tag{2.2}$$

μ_m and Σ_m are the mean vector and covariance matrix of the m^{th} component respectively [75]. The component weights $w_m > 0$ must satisfy $\sum_{m=1}^{M} w_m = 1$. In practice, for reasons of data requirement and computation, covariance matrices are usually diagonal [78].

Training a GMM involves finding the parameters $\lambda = \{w_m, \mu_m, \Sigma_m\}_{m=1}^{M}$ given a training sample $x = \{x_1, x_2, ...x_T\}$. The Log-likelihood (LL) of x with respect to λ is given by:

$$LL = \frac{1}{T}\sum_{t=1}^{T} log_p(x_t|\lambda) \tag{2.3}$$

The higher the value of LL, the stronger the indication that x originates from the GMM λ. Maximun Likelihood (ML) estimation is an approach to maximize the likelihood of the model with respect to the given data, and can be achieved with the iterative Expectation Maximization (EM) algorithm [13].

2.9.3 Testing phase of GMM classifiers

Having extracted the features and trained a GMM for each class, test data is then used to evaluate the system. In the recognition mode, the 'match score' between feature vector $X =$

$\{x_1, x_2, ...x_t\}$ and a model λ_s is given by the *LL*. The *LL* for each λ_s is used to predict which model the test vector belongs to.

Speech processing techniques have been successfully exploited in the analysis of calls and songs from bird populations. Techniques used for years in speaker classification and speech recognition, have been shown to have the capability to be adapted to the task of bird species recognition.

2.9.4 Classification using HMMs

HMMs have been used for speech recognition since the late 70's, but after Lawrence Rabiner published [126, 125] in the late 80's, the popularity of HMMs increased [122].

While the use of HMMs for automatic speech recognition is well known, HMMs have also been applied to other acoustic applications. HMMs were used for acoustic event detection in [10]. HMMs were used for bird classification in [165, 67, 48]. HMMs were again used as a baseline for bird phrase classification in [71]. Songscope [1, 173] uses HMMs for bird recognition and likewise Wei Chu [31] used HMMs to detect Robin syllable boundaries.

There are three basic problems when dealing with HMMs:

1. Evaluation: Given a HMM model, λ, and a sequence of observations, $X = x_1, x_2, ..., x_T$, what is the probability that the observations are generated by the model, $p(X|\lambda)$?

2. Decoding: Given a HMM model, λ, and a sequence of observations, $X = x_1, x_2, ..., x_T$, what is the most likely state sequence in the model that produced the observation? Viterbi is commonly used to decode this.

3. Learning: Given a HMM model, λ, and a sequence of observations, $X = x_1, x_2, ..., x_T$, how should the model parameters, $\lambda = \{A, B, \pi\}$, be updated in order to maximize $p(X|\lambda)$?

A GMM is most widely used distribution within each state of a HMM. HMMs are useful to model variation in features and temporal evolution. HMMs can be used to model discrete observations (colors within a particular urn, example in [126]) or continuous observations (more common for audio processing). When dealing with continuous observation densities, every state contains a GMM with weight w_m, mean μ_m, and covariance Σ_m, with mixture number $m = 1, ..., M$. The state transition probabilities, A, determine how likely it is to move from one state to the next during a state sequence. Each state j has a probability distribution b_j, with B denoting probability distributions for all states (see Eq. 2.4 and Eq. 2.5).

$$b_j = \{b_j(k)\}, \qquad 1 \leq k \leq M \tag{2.4}$$

$$B = \{b_j\}, \qquad 1 \leq j \leq N \tag{2.5}$$

where j is state number, N is the number of states, k is the mixture number, M is the number of mixtures per state, B is the model observation probability distribution and b_j is the probability distribution in state j.

The initial state distribution is π. π is used to determine the first state of an observation sequence. In a classification problem, each class will have a model λ (see Eq. 2.6).

$$\lambda = \{A, B, \pi\} \tag{2.6}$$

More details on HMM theory is given in Appendix B.

2.10 Collaboration of Engineering and Zoology

Decisions on avian taxonomy of species lists are frequently being reviewed [160]. Reassignment of species limits may be necessary in poorly studied regions where original species distinctions were not made scientifically. Revisions may need to be made in cases where a population of species has evolved or diversified enough to be treated as a new unique species. Bird vocalizations are an important facet of these review processes. Signal processing can help zoologists measure acoustic difference automatically. Section 2.8 introduced theories and methods used by zoologists when analyzing bird species limits, vocal difference and the study of current species taxonomic classifications.

In chapter 5 a taxonomic measurement system, used by Tobias et al. in [160], which uses acoustic, morphological, plumage, ecological, behavioral and geographical evidence, will be described in detail. As voice was one of the features used by Tobias et al., the vocal comparison system inspired experiments investigated in chapter 5. Acoustic evidence in [160] was extracted by analyzing spectrograms and using on-screen cursors to measure the features. This method is exhaustive and can be prone to subjective selection. Work in chapter 5 involves trials to automatically extract the vocal features used in [160], on a similar dataset. Automatic feature extraction will greatly benefit ornithologists in their future studies.

A different approach to bird song population difference is described in chapter 3. To echo the message of previous Sections, measures of difference (or similarity) in vocalizations of bird populations are of great importance to the ornithology community. In speech processing, knowing how close or diverse dialects are provides useful information to predict system performance. Dialect variations of a language have a severe impact on the performance of speech systems. Therefore, knowing how close or diverse dialects are in a given language space provides useful information to predict, or improve, system performance when there is a mismatch between train and test data [102]. In [102,103], Mehrabani et al. investigated a method of dialect distance assessment based on pitch patterns modeled progressively from pitch contour. This work inspired the theory that bird population difference could be measured using a similar pitch contour comparison. This led to experiments presented in chapter 3. A difference measure was computed by comparing pitch codebooks of two bird populations. Populations classified as the same species

should have a small difference value where as populations said to be different species should have a larger difference value.

This thesis focuses on a number of key research questions like can pitch be used to quantify bird vocalization difference between population pairs? To help answer this, technologies used by speech processing when measuring speaker dialect difference are drawn upon. A method used by zoologists is also examined in chapter 5. Another research question is, can pitch tracking of bird vocalizations improve in accuracy? Current pitch tracking techniques for bird song are taken from speaker pitch detection. Work in this thesis investigates an optimization approach for bird pitch tracking presented in chapter 4. Segmentation of long field recordings with accurate song boundaries and issues surrounding calculating the number of notes in a song automatically, is another theme central to this thesis. Can this segmentation analysis be done appropriately? This motif is explored in chapter 6. Probably the most important question investigated in this thesis is, how do zoologists review classification limits? Accurately extracting pitch contours and song durations are necessary for their bird analyses, when reviewing these limits. All these key questions will be investigated in the Chapters that follow.

The rest of the thesis is laid out as follows. Chapter 3 describes work using pitch to compare populations of birds inspired by speech processing. Optimization of pitch tracking for birds is presented in chapter 4, which again draws on technologies used in signal processing. An acoustic comparison system for birds used by Tobias et al. [160], which uses temporal and spectral features, is discussed in Chapter 5. This introduces a system from the zoology community. Experiments in Chapter 5 apply signal processing tools to help automate this system. This brings the field of engineering and zoology together. Experiments in chapter 5 require accurate song boundaries to calculate song duration. Measures also need to be taken to ensure that pitch is only extracted from bouts of target bird song. Currently hand annotated segmentation textgrid files of bird song boundaries are required. Automating this segmentation process is undertaken in chapter 6.

3

Quantifying difference in vocalizations of bird populations

In speech processing, researchers are interested in the difference (or similarity) of different speech dialects of a single language. Work by Mehrabani et al. in [102, 103], used statistical representations of pitch to derive a measure of the difference between different dialects of Arabic. Given the success of speech processing methods applied to bird species classification, chapter 3 presents work on developing a measure of bird call difference. The first experiment in this thesis exploits the method presented in [102, 103], extending it to the domain of bird calls and short songs. Work here reports the first stage of work in developing a measure of call difference for birds to aid ornithologists in their work. The measure is applied to bird populations with calls that are considered very similar, very different and between these two extremes. Initial results are very promising, with the behavior of the metric consistent with accepted levels of difference for the populations tested to date. A portion of this work was presented by O'Reilly et al. in [116].

This chapter presents work as follows. First the adaption of method used in [102, 103] for use with bird vocalizations difference is presented. Then a description of how Fundamental Frequency ($\boldsymbol{F_0}$) or pitch contours extracted from bird vocalizations are analyzed using this method is presented. Next the extension of the codebook from a 3x3 to a 5x5 codebook, which allows a more accurate representation of the pitch contour structure, is examined. Work to this point was presented in [116] at INTERSPEECH 2015. Following that publication, inspection of preliminary work showed that pitch estimates for bird calls were prone to errors. This highlighted the need for improved pitch tracking of birds and YIN-bird was developed (discussed in Chapter 4).

Difference scores using the improved pitch tracker were then calculated, with the new results also presented in this chapter for continuity. The new pitch tracker, YIN-bird is fully presented in chapter 4. Finally difference measures are reported when replacing the linear codebooks with VQ codebooks. An analysis of how sensitive difference measures are to data size is also given.

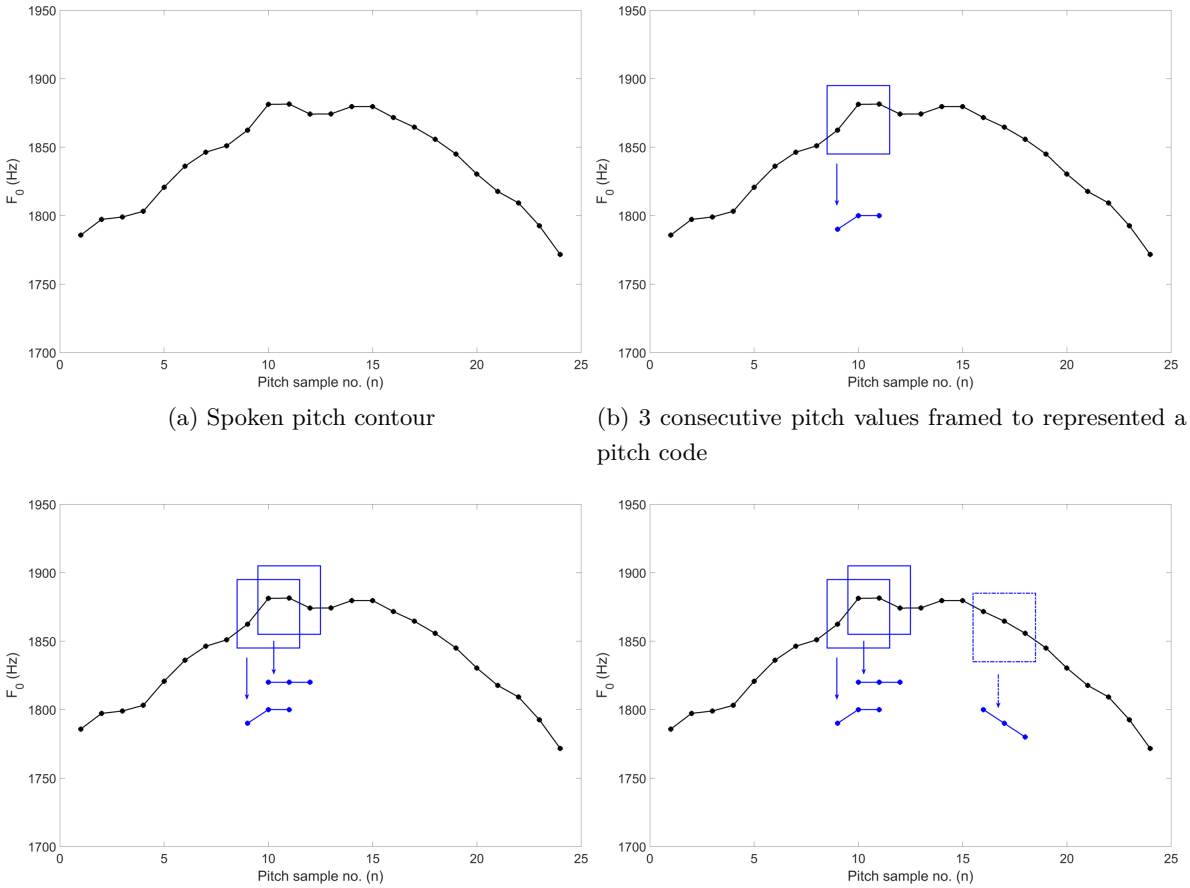
3.1 Method - Spoken dialect pitch contour difference

Pitch contours provide cues for tasks such as language and dialect identification. Lin and Wang in [30] used Legendre polynomial coefficients to represent segments of pitch contours for language identification. Biadsy and Hirschberg [12] demonstrated that four Arabic dialects exhibited significant differences from one another in terms of prosodic structure, including pitch range, register, and pitch dynamics. After human perception tests indicated that prosodic cues, including pitch movements, can be used to identify one accent from another, Mehrabani et al., in [102, 103], exploited this in a system for measuring difference between dialects via differences in pitch contour microstructures. As birds production and perception is similar to humans [7], this approach may be applicable to species differentiation. Figure 3.1a shows pitch values from a speech clip. An example of a pitch analysis window is shown in Figure 3.1b by a blue box surrounding three $\boldsymbol{F_0}$ samples. Every pitch analysis window contained three consecutive non-zero $\boldsymbol{F_0}$ values, from a speech recordings, with an overlap of two samples between adjacent windows. An example of this overlap can be seen in Figure 3.1c where two adjacent windows shared two $\boldsymbol{F_0}$ values. Each window [F_{01}, F_{02}, F_{03}], was then transformed into a 2D vector [$(F_{02} - F_{01})$, $(F_{03} - F_{02})$]. [$(F_{02} - F_{01})$, $(F_{03} - F_{02})$] is equal to $[\Delta F_{01}, \Delta F_{02}]$. This 2D vector was converted into a code number using a codebook. The 3x3 codebook can be seen in Figure 3.2a. This work was expanded to a 5x5 codebook here and 5x5 codebook is shown in Figure 3.2b. The 3x3 codebook has nine codes determined by $[\Delta F_{01}, \Delta F_{02}]$'s relationship to $\pm$ threshold 1. This pitch analysis procedure was applied to recordings from different dialects, D_i (i is the dialect number). Each dialect being compared, i.e. D_1 and D_2, will have a sequence of pitch codes. In Figure 3.1b, three consecutive pitch estimates are grouped together and these correspond to code 6 from the 3x3 codebook (shown in Figure 3.2a). Code 6 is an upsloped followed by a flat pitch contour. Figure 3.1c includes a second example which corresponds to code 5 from the 3x3 codebook. Figure 3.1d includes a third example which corresponds to code 7 from the 3x3 codebook. Codes 5 and 7 describe a flat contour and a downsloped contour respectively.

A more comprehensive description of the dialect distance extraction is available in [102, 103] but a brief summary is provided here to aid in understanding the current work.

After extracting $\boldsymbol{F_0}$ and counting the occurrences of each pitch code, the conditional probabilities were calculated using Eq. 3.1 for each class and stored in a [9x9] matrix.

$$P(w_i|w_j) = \frac{C(w_j, w_i)}{C(w_j)}, i, j = 1, 2, ..., 9 \tag{3.1}$$



(a) Spoken pitch contour

(b) 3 consecutive pitch values framed to represented a pitch code

(c) Second example of pitch code (a flat contour)

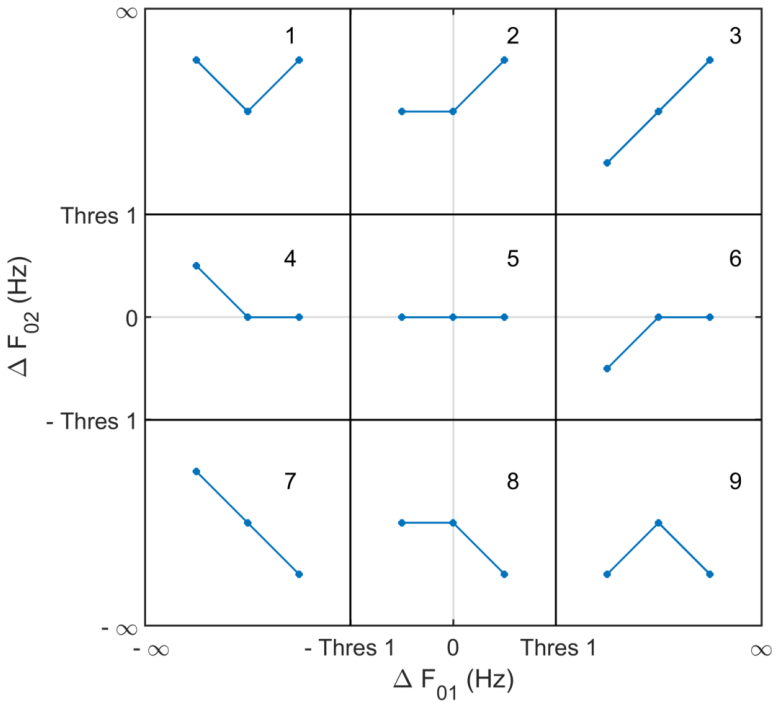
(d) Third example of pitch code (decending contour)

Figure 3.1: Description of a spoken pitch contour example transformed into codes from the 3x3 codebook based on work from [102, 103]. (a) Spoken pitch contour. (b) Three consecutive pitch values framed to represented a pitch code. (c) Second example of pitch code (a flat contour). (d) Third example of pitch code (descending contour).

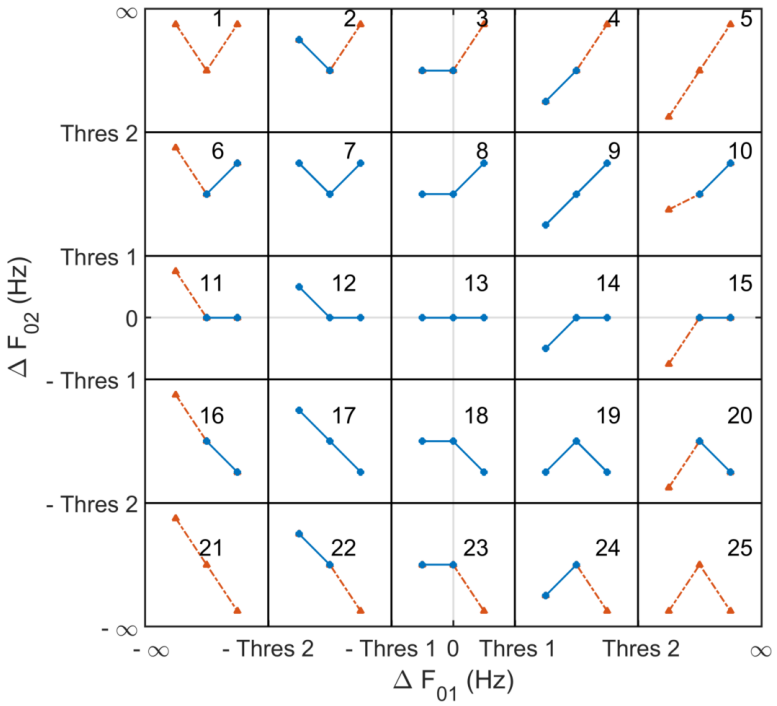
In Eq. 3.1, $C(w_j, w_i)$ is the number of transitions from code w_j to code w_i, $C(w_j)$ is the number of occurrences of code w_j and $P(w_i|w_j)$ is the probability of transition. Next the Kullback Leibler (KL) divergence was used to derive a difference metric between the different populations. The KL divergence or relative entropy is a non-commutative measure of similarity/dissimilarity between distributions or statistical models [102]. The KL method was first published in [83].

$$d(D_1, D_2) = \sum_{i=1}^{n} \sum_{j=1}^{n} P_1(i,j) log \frac{P_1(i,j)}{P_2(i,j)} \tag{3.2}$$

In Eq. 3.2, $d(D_1, D_2)$ is the KL divergence of D_2 from D_1, n is the codebook size (e.g. 9 for 3x3 or 25 for 5x5), $P_1(i,j)$ and $P_2(i,j)$ are discrete pitch pattern distributions for dialect 1 and dialect 2 respectively. $d(D_1, D_2)$ and $d(D_2, D_1)$ are not necessarily equal, therefore the average



(a) 3x3 pitch codebook



(b) 5x5 pitch codebook

Figure 3.2: Pitch contour codebook. (a) 3x3. (b) 5x5. Orange line indicates pitch slope greater than threshold 2, while blue represents a flat contour or rise/fall in pitch between threshold 1 and threshold 2.

of the two differences was taken.

Work in [102] focused on dialect difference using pitch pattern bi-gram methods. While work in [103] also explored other metrics including log likelihood scores distribution, pitch pattern uni-gram and energy pattern bi-gram methods, only the bi-gram pitch pattern method was used in this thesis, which was found to be a robust indicator of dialect difference [103].

3.2 Bird Data

Four different data sets are used for experiments here. Set 1 contains two subspecies thus it is anticipated calls of the two groups are similar. Set 2 contains two subspecies from three locations. Vocalizations from two locations are from one subspecies while vocalizations from the third location are a different subspecies. Zoologists believe there is hard evidence that this third population may in fact be a species in it's own right and may be reclassified as a separate species following genetic analysis. Difference in the calls of this population will add weight to its reclassification, hence acoustic analysis of this subset of data is of great importance. Set 3 contains two different passerine species from different families, which means they are very distantly related and hence should have very different vocalizations. Finally set 4 contains three sets of recordings from different species of Mexican antbird. Each population of antbird is a different species. The amount of data in each set is summarized in Table 3.1. Pictures of the different bird populations can be seen in Figure 3.4.

3.2.1 Indonesian island populations - Set 1 & 2

Recordings of Black-naped Oriole (BNO) and Olive-backed Sunbird (OBSB) calls were collected by TCD Zoology Department during field trips to Indonesia. Harte et al. used this data to build GMM classifiers in [59]. All recordings were made with a Zoom H2 Handy Recorder [178] using a separate microphone (Sennheiser ME62 Omni-Directional Condenser Microphone Capsule [141]). The microphone was mounted in a parabolic reflector, specifically a Telinga Reflector with a 1mm flexible dish [158]. Bird recordings were stored as MP3 files at 128 kbps as a compromise between quality and storage limitations when in the field. Recordings were taken during the dawn and dusk choruses, when the greatest number of birds were vocalizing. Careful procedures for recording were in place to ensure the same individual bird was not repeatedly recorded. Recording distances varied from a few meters to as much as 100 meters (which presented some reverberation issues). Adobe Audition was used to segment each bird recording into individual calls from that bird. Although each file contained one call, the start and end of the files were not flush to the call timestamp. Many files are longer than the actually call. The average file length is displayed in Table 3.1 but this is slightly longer than the average call duration. Some pre-processing was performed to minimize any spurious noise such as background cars, crickets, hum, or running water that had obscured recordings. The recording process description was

Table 3.1: Bird data for sets 1-4.

Black-naped Oriole (Set 1)	**No. of Calls**	**No. of Birds**	**Total Length (s)**	**Mean File Duration (s)**
Buton	110	15	124.14	1.13
Wangi-Wangi	77	10	87.52	1.14
Olive-backed Sunbird (Set 2)	**No. of Calls**	**No. of Birds**	**Total Length (s)**	**Mean File Duration (s)**
Buton	517	11	152.66	1.13
Sulawesi	218	7	57.92	0.27
Wangi-Wangi	268	10	168.06	0.63
Species (Set 3)	**No. of Syllables Syllables**	**No. of Recordings**	**Total Length (s)**	**Mean File Duration (s)**
Alder Flycatcher	515	33*	191.53	0.37
Barn Swallow	575	77*	282.93	0.49
Mexican antbirds (Set 4)	**No. of Calls**	**No. of Birds**	**Total Length (s)**	**Mean File Duration (s)**
Dusky Antbird	1332	68	2100.00	1.57
Dot-winged Antwren	236	25	539.99	2.29
Mexican Anthrush	913	79	1903.52	2.08

* Not known if each recording contains a unique bird subject

presented in [59].

The BNO found on the Wakatobi islands (including Wangi-Wangi (WW)), most likely colonized from the regions of Sulawesi (SU) and Buton (BU). A map of the Indonesian territory where the recordings were made is given in Figure 3.3. The BU and WW populations of BNO are described as Set 1 here. The BU population of BNO is recognized as *Oriolus chinensis celebensis*. The WW population is a different subspecies recognized as *Oriolus chinensis boneratensis*. As subspecies of a single species, it is predicted their calls would be similar [59]. An image of a BNO individual from BU is shown in Figure 3.4a.

Three island populations of OBSB are studied here as Set 2, with calls from BU, SU and WW. The OBSB from mainland SU and BU are officially recognized as subspecies *Cinnyris juglaris plateni*. The birds found on WW, are officially recognized as subspecies *Cinnyris jugularis infrenatus*, but there is strong evidence that these Wakatobi birds may be a separate species. Whilst the reclassification of any bird is a multifaceted task, the individuality of their song is as important as morphological and plumage differences [59]. If the WW bird vocalizations show significant and repeatable differences from the BU and SU OBSB populations, this will add significant weight to its classification as a different species. An image of an OBSB individual



Figure 3.3: Map of Indonesian region where Sets 1 & 2 were recorded [59].

from SU is shown in Figure 3.4b.

3.2.2 Species - Alder Flycatcher & Barn Swallow - Set 3

Set 3 contains two species from different families: the Alder Flycatcher (AF) and Barn Swallow (BS). This data was used to represent a pair of birds that are very different. The species classifier in [57] by Graciarena et al., showed both species to have a low Equal Error Rate (EER). This is consistent with their taxonomic difference. Using an online database 'http://www.xeno-canto.org/' [166], 33 AF recordings and 77 BS recordings, labeled as 'A' quality, were acquired. These recordings required background bird vocalization removal and segmentation. The segmentation was done in Matlab using a modified version of 'Harma Syllable Segmentation' (code by Michael Lindermuth [91]). After pre-processing, there were approximately 500 syllables for each species. The recordings here are thus not from the same dataset used in [57]. Images of AF and BS individuals are shown in Figures 3.4c and 3.4d.

3.2.3 Mexican antbirds - Set 4

Chu et al. used an antbird corpus in [172], which was also used by Trifa et al. in [165]. The data was collected from a Mexican rainforest by researchers from the Ecology and Evolutionary Biology department at UCLA [165, 167]. The antbirds shared the same territory located at the Estacion Chajul in the Reserva de la Biosfera Monte Azules in Chiapas, Mexico [165].

These recordings are more complex than the calls of sets 1-3. The Mexican Anthrush (MAT) is part of the *Formicariidae* family while the Dot-winged Antwren (DWA) and the DAB are part of the *Thamnophilidae* family. For ease, they will be referred to as MAT, DWA and DAB

Table 3.2: Various bird classifiers

Set	Data	Classification Rate (%)	No of. Mixtures
1	BNO(BU vs. WW	66.67	8
2	OBSB (BUSU vs. WW	93.82	20
3	AF vs. BS	93.54	20
4	antbirds (DAB, DWA, MAT)	98.12	256

respectively. It is predicted that the MAT's vocalizations will be more different than the DWA and DAB's, but it should still be possible to separate DAB and DWA calls as they are different species. The analysis of these birds is compared with other bird populations later in this Chapter. An image of a DAB individual is given in Figure 3.4e.

3.2.4 GMM Classification

To give an indication of how separable these birds calls are, each set of data was used to build a GMM classifier (one system for each set). The features used for the bird classifiers in this chapter were 12 MFCCs, log energy, delta and double delta coefficients. These features were extracted using a 25 ms hamming window with a frame rate of 10 ms. This window size was reported as optimal for birds in [56]. GMMs were trained with varying numbers of mixtures, dictated by the amount of data available for each set. The same data listed in Table 3.1 was used. Results of the GMM classifier systems are shown in Table 3.2. Due to lack of data, cross-validation was used for BNO, OBSB and AF/BS classifiers. When building the Mexican antbird three class problem (DAB, DWA, MAT), a 70% train 30% test split was used. The BNO has a low classification accuracy which is expected when the classes are the same species. The antbirds and AF versus (vs.) BS classification accuracies are high because the data sets contain different species. The OBSB classifier had two models. One model was trained using BU and SU OBSB recordings as they are the same sub species and the second model was trained using WW OBSB as this population is not only a different subspecies, but it is suspected to have diversified enough to become a species in it's own right. OBSB are correctly recognized 93.82% of the time, even though the populations are currently classed as different subspecies of the same species. The classification accuracy of the two class BNO classifier is considerably lower at 66.67% which is only slightly better than chance for a two class problem. This suggests that the BU and WW subspecies of BNO are similar which supports the current classification that they are different subspecies of the same species.



(a) BNO



(b) OBSB



(c) AF



(d) BS



(e) DAB

Figure 3.4: Images of idividual birds from (a) BNO [59], (b) OBSB [59], (c) AF [39], (d) BS [39] & (e) DAB [39] species.

3.3 Dialect difference for bird vocalizations

The work presented in [102,103] showed how pitch contours can be used to derive a dialect difference measure. This prompted the question can similar ideas be applied to bird calls to measure population difference? Work in this chapter presents experiments which attempt to answer this question. There are fundamental differences between human speech and bird vocalizations as discussed in Section 2.1.3. These differences allow birds to rapidly change pitch faster than humans [25]. Bird vocalizations tend to have higher $\boldsymbol{F_0}$ and are capable of changing pitch much more rapidly. To account for this, modifications needed to be made to the system in [102, 103]. Here instead of dealing with dialect 1 and 2, the difference of species/population 1 (S_1) and species/population 2 (S_2) is measured. An overview of the proposed modified difference system inspired by [102, 103] is given in Figure 3.5.

The first stage of the system involves extracting the $\boldsymbol{F_0}$. In [102, 103] RAPT (taken from Voicebox for Matlab [20]) was used to extract the $\boldsymbol{F_0}$ as they were dealing with human speech. YIN [42] is more accurate when dealing with higher pitch signals like bird calls. YIN is used by Sound Analysis Pro (SAP) for $\boldsymbol{F_0}$ extraction and is discussed by Mandelblat-Cerf in [95].

Window sizes of 25, 30, 10 and 10 ms were selected for the BNO, OBSB, Mexican antbirds and AF/BS respectively. The frame rate was 5 ms for all sets. These values were chosen experimentally by observing the best fit pitch contour line over the spectrograms. Longer windows give a better pitch estimate but shorter windows are required for rapidly modulating bird vocalizations.

After $\boldsymbol{F_0}$ extraction the difference between adjacent samples, ΔF_0, was used to calculate the thresholds used by the 3x3 codebook. Each set of populations had a 3x3 codebook (introduced in Section 3.1 and displayed in Figure 3.2). Threshold 1 was chosen so that approximately 70% of ΔF_0 was within $\pm$ Threshold 1 for each set of populations. In [102, 103], Mehrabani et al. used a threshold of $\pm$ 3 Hz. After analyzing two hours of male speech, approximately 70% of the $\boldsymbol{F_0}$ was found to be within $\pm$ 3 Hz, hence 70% was used as a target for each species' threshold as a reasonable first choice. A data set specific threshold was established as 20, 50, 50 and 150 Hz for the BNO, OBSB, antbirds and AF/BS combination, respectively (summarized in Table 3.3). These simple modifications were used to adapt the original system from [102,103] to form a new system used here to assess bird call difference between populations.

3.3.1 Difference measure with 3x3 codebook

The first pitch difference experiment used the modified system with a 3x3 codebook. Table 3.4 shows the difference score results with 3x3 codebooks for Sets 1-4. Table 3.4 also contains a description of how different each bird population pair is based on current taxonomy and ornithologist's knowledge. This knowledge was provided during private communication with David J. Kelly from the department of zoology, TCD. The two BNO populations yield a small difference of 0.0455, which for populations with very similar calls should be small relative to

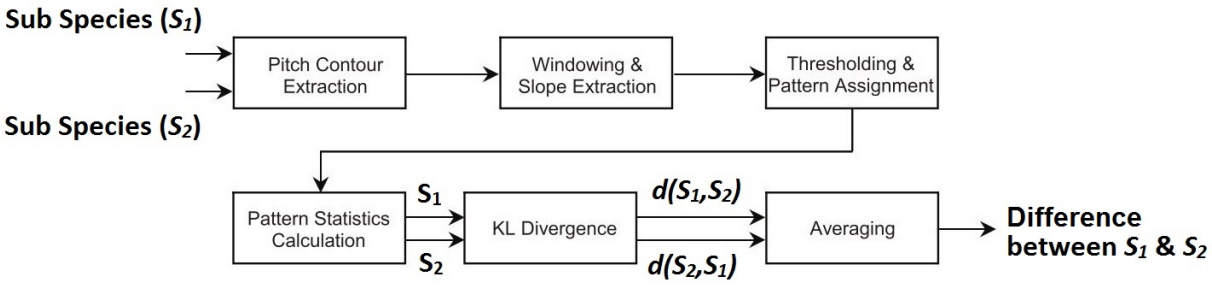


Figure 3.5: Block diagram of bird pitch difference system [102, 103].

Table 3.3: Thresholds for bird call difference system

Set	Data	Threshold 1 (Hz)	Threshold 2 (Hz)
1	BNO	20	40
2	OBSB	65	120
3	AF/BS	150	400
4	Antbirds	50	100

populations with very different calls. The differences reported for the MAT compared to both the DWA and DAB yield the largest difference of 0.8156 and 0.822 respectively (see Table 3.3). This is consistent with the expectation that the vocalizations from these birds are very distinct. The DAB and DWA comparison and AF versus BS yield comparable differences of 0.2278 and 0.3588. These measures are smaller, but are still from quite distinct populations. Thus measures of between 0.35 and 0.82 all correspond to identifiable difference in call. The measure of the OBSB at 0.5155 falls between these extremes. The large difference for the OBSB supports the hypothesis that the subspecies in question has diversified, supporting the suggestion that the WW population be reclassified as a separate species. These initial results are very promising but this measure of difference still requires refinement. As a first step, it was decided to explore whether allowing a more detailed representation of the pitch slopes for a given species was useful. Thus the 3x3 codebook was extended to a 5x5 codebook by introducing a second threshold. The increase to two thresholds allows the pitch to be described as increasing dramatically, increasing, stable, decreasing or decreasing dramatically. This is attractive given the more complex pitch structure of some bird vocalizations. For example, the Mexican antbirds vocalizations are more complex than the shorter less melodic calls of the BNO and OBSB.

Table 3.4: Bird call difference, 3x3 codebook.

Set	Species	3x3	Description
1	BNO(BU vs. WW)	0.0455	V. similar, different subspecies
2	OBSB (BUSU vs. WW)	0.5155	WW speciation in dispute, different subspecies
3	AF vs. BS	0.3588	V. different, different species
4a	DAB vs. DWA	0.2278	Different, same family, different species
4b	MAT vs. DWA	0.8156	V. different, different family, different species
4c	MAT vs. DAB	0.822	V. different, different family, different species

Table 3.5: Bird call difference, 5x5 codebook.

Set	Species	5x5	Description
1	BNO(BU vs. WW)	1.0019	V. similar, different subspecies
2	OBSB (BUSU vs. WW)	2.5724	WW speciation in dispute, different subspecies
3	AF vs. BS	3.5661	V. different, different species
4a	DAB vs. DWA	1.5238	Different, same family, different species
4b	MAT vs. DWA	2.5971	V. different, different family, different species
4c	MAT vs. DAB	2.9953	V. different, different family, different species

3.3.2 Extension to a 5x5 codebook

The 3x3 codebook was extended to a 5x5 codebook by introducing a second threshold. The second threshold again was data set specific. 40, 100, 100 and 400 Hz was chosen experimentally as Threshold 2 for the BNO, OBSB, antbirds and AF/BS combination respectively (again see Table 3.3). Threshold 2 was chosen so that approximately 90% of that species' ΔF_{01} and ΔF_{02} were within $\pm$ the selected value (Threshold 2). The increased number of codes gives more information about the bird's pitch changes.

By increasing the codebook to a 5x5 all difference values will increase due to the KL formulation. Thus comparisons with Table 3.4 are not informative but patterns should still emerge in Table 3.5. The behavior of the measure remains consistent for the antbird populations with measures between 1.52 and 3.00 for these diverse songs. Similarly the BNO measure is much smaller at 1.002. The OBSB results at 2.5724 falls firmly into the range of different. The AF versus BS measure at 3.5661 is now the highest of the measures, supporting the diversity of call between the two species. Once more, the difference measure for the OBSB is of a similar order to measures for song recognized as quite diverse. This result is consistent with stating that the OBSB population for WW has a very distinct call compared to the subspecies of BU and SU.

3.4 Accuracy of pitch contours

An accurate difference measure using the system described in Section 3.3 is reliant on accurate F_0 or pitch estimates of the bird data. Pitch contour analysis on speech does not require prior knowledge, manual segmentation or pre-processing. Pitch tracking and voice activity detection have been optimized for performance on speech in controlled environments [42, 152]. The majority of bird recordings are made in the bird's natural habitat. Many factors can influence F_0 estimation. A common difficulty with bird song analysis is data is often made up of degraded versions of the signals sent by the birds. Thick vegetation, like in the canopy of a rainforest, causes signal scattering. High frequency sounds are affected by small obstacles such as leaves while low frequencies are unaffected. This results in a low pass filtered signal arriving at the microphone. Other forms of degradation include discrete echoes, sound tails and cancellations due to the ground effect. Even a bird turning its head while vocalizing can influence the amplitude distribution of the harmonics being picked up by a microphone. Signal degradation was discussed in depth by Marler et al. in 'Nature's Music: The Science of Birdsong' [96]. Another issue with recording birds in their natural habitat is that one cannot control what birds are vocalizing at a given time. Recordings usually contain a target individual from the species under analysis and multiple conspecific or heterospecific individuals.

While efforts were made to minimize the non-target vocalizations, inevitability some will be present in the data. Pre-processing of BNO and OBSB data removed non-target audio from parts of the spectrogram not containing target information. This was done manually using Adobe Audition. Even with this processing, BNO and OBSB data was still affected by noisy and reverberant conditions. AF and BS data was segmented using an amplitude threshold with the assumption that high amplitude information was from the target (using code at [91]). As AF and BS data was large and not under reclassification, the pre-processing was not given the same attention as the BNO and OBSB. Finally the Mexican antbird data had already been pre-processed when it was shared. Details of pre-processing is given by Wei Chu in [31].

Even with very clean data, F_0 estimates can still be subject to errors. Extended frequency sweeps through multiple octaves and rapid pitch modulations are just some of the difficulties encountered when estimating the pitch of bird song. YIN, which is based on the autocorrelation method (but using average squared difference instead), is very effective for pitch tracking, but some peaks suffer ambiguity, which leads to octave error or estimates too low in frequency [87]. This highlighted the need to improve pitch tracking performance for birds. YIN-bird was developed and is discussed at length in chapter 4.

3.5 Comparison of F_0 estimation using YIN-bird for bird call difference

A comparison of bird call pitch extraction using YIN and YIN-bird, along with difference scores achieved using either of the two will now be discussed. For the most part YIN tracks F_0 of bird vocalizations quite well. The use of YIN-bird is not always necessary to acquire accurate F_0 contours as shown in Figure 3.6. The figures in this section compare pitch tracking using YIN on the left hand side with using YIN-bird on the right hand side. F_0 estimates from both match the spectrograms quite well in the examples in Figure 3.6.

Unfortunately this YIN performance is not guaranteed for every example. Figure 3.7 shows examples of octave errors or errors too low when using YIN. The right hand side of Figure 3.7 shows these errors corrected.

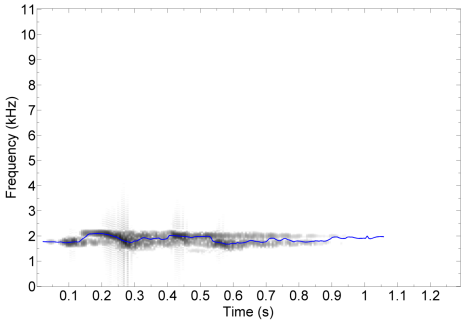
Some calls are very short and thus frequency tracks are hard to make out when viewing the spectrogram. Short calls are particularly prone to errors with YIN but YIN-bird is better able to deal with the task (see Figure 3.8).

A major issue with pitch tracking of birds recorded in crowded locations is non-target corruption of the acoustic space by other birds. While experts can easily identify different targets by visual inspection of spectrograms or listening to recordings, F_0 systems will try and track the pitch of anything it can. Both YIN and YIN-bird track the F_0 of target and non-target birds in Figure 3.9. While YIN-bird removes octave errors it still tracks the pitch of some non-target information.

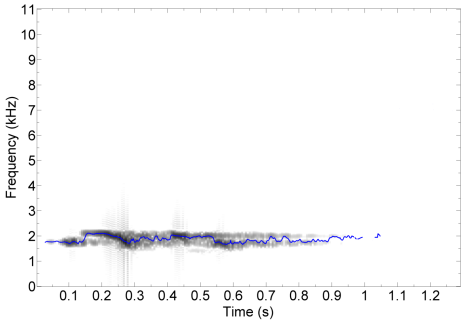
Another issue with pitch contours used in this chapter are reverberation tails. It is inevitable when recording targets from up to 100 meters, that reverb tails will be captured too. Both YIN and YIN-bird track F_0 of reverb tails. Extreme examples if this can be seen in Figure 3.10.

At first, noise was thought to be an issue which influenced errors, but upon inspection of signals with noise and signals with noise removed, noise appears to have little influence on octave errors. Some examples are shown in Figure 3.11.

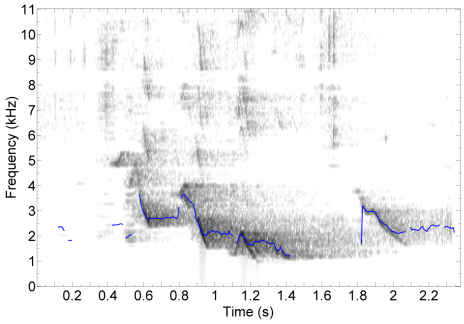
For ideal pitch contour analysis, recordings with one individual, little reverb, little noise and labeled start and end times are desired. Unfortunately recordings often do not fit these criteria. YIN-bird helps with removing octave errors and prevents tracking low frequency non-target events. It does not help with the tracking of reverb tails and is something to be aware of during the analysis. Dereverberation techniques have not been investigated on the data here.



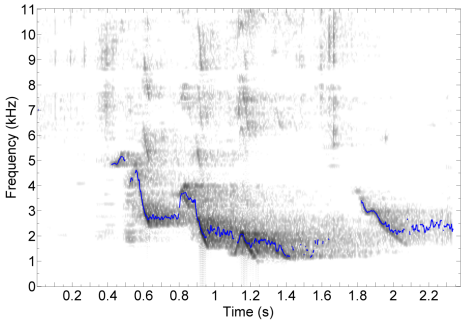
(a) YIN - 03_Yin_buoriole_01_0003



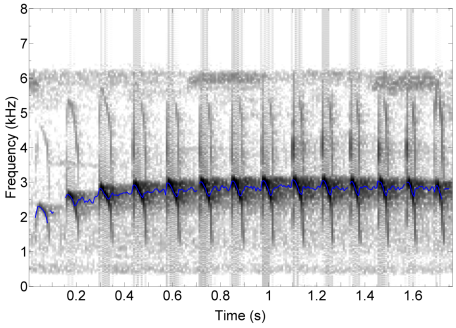
(b) YIN-bird



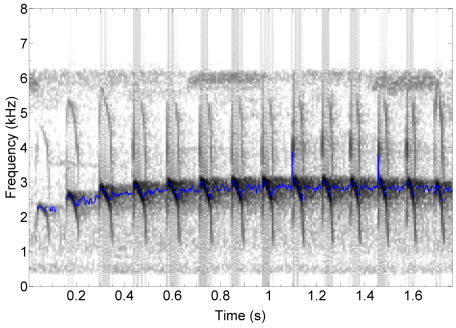
(c) YIN - 04_Yin_wworiole_05_0001



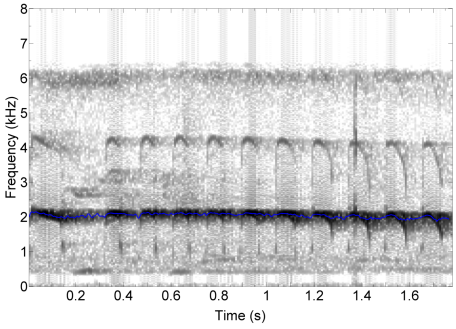
(d) YIN-bird



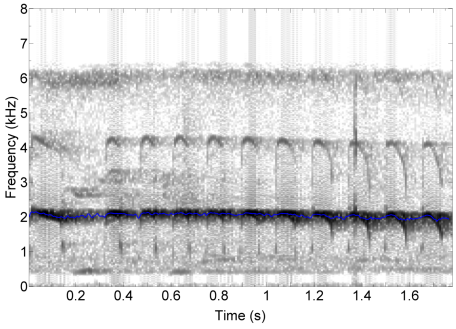
(e) YIN - 21_Yin_D020306C131009



(f) YIN-bird

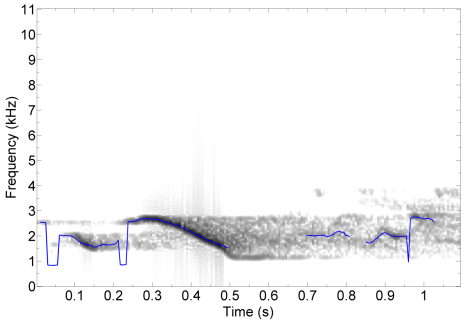


(g) YIN - 23_Yin_M060605K060018

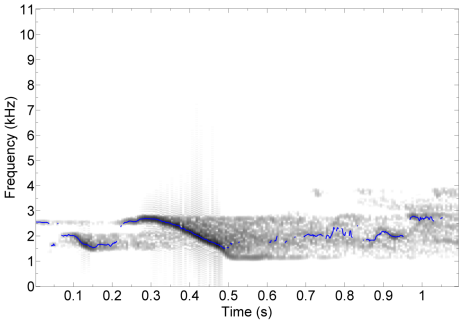


(h) YIN-bird

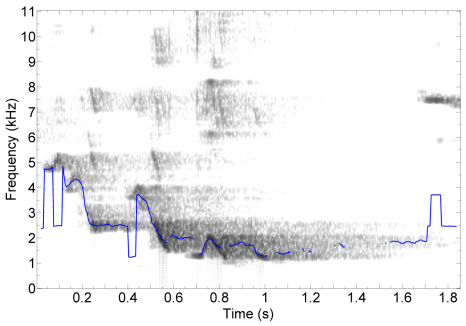
Figure 3.6: Examples of YIN F_0 estimation with good performance. Figures on the left hand side show YIN F_0 contours, while plots on the right hand side show F_0 estimation using YIN-bird with the same examples. F_0 estimates are superimposed in blue.



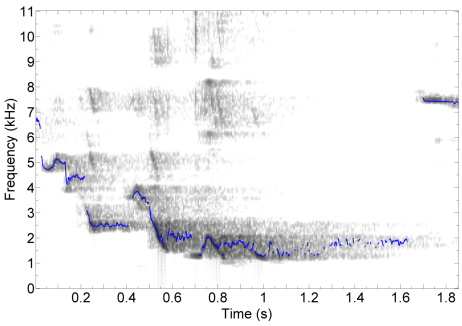
(a) YIN - 01_Yin_buoriole_06_0001



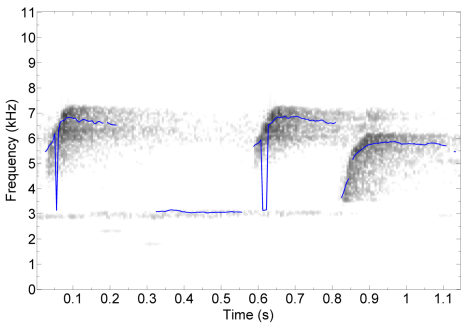
(b) YIN-bird



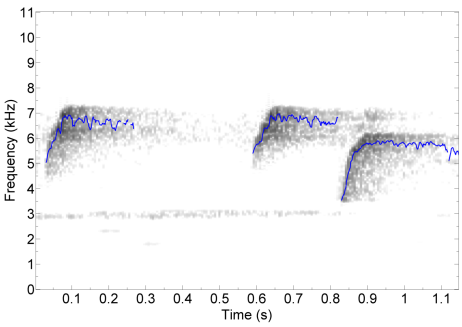
(c) YIN - 05_Yin_wworiole_05_0002



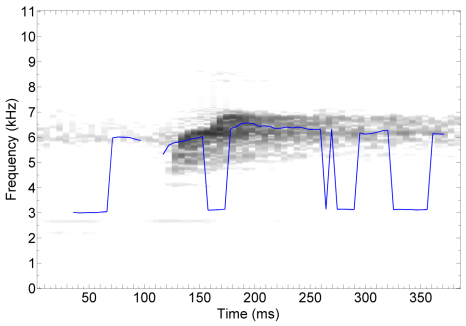
(d) YIN-bird



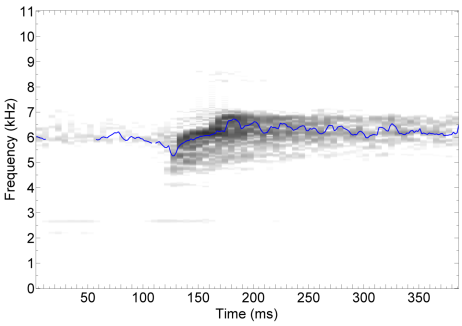
(e) YIN - 06_Yin_buobsb_01_0002



(f) YIN-bird

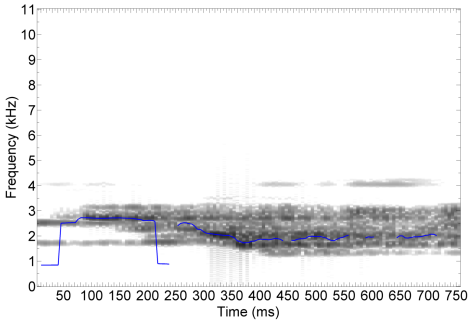


(g) YIN - 08_Yin_buobsb_01_0009

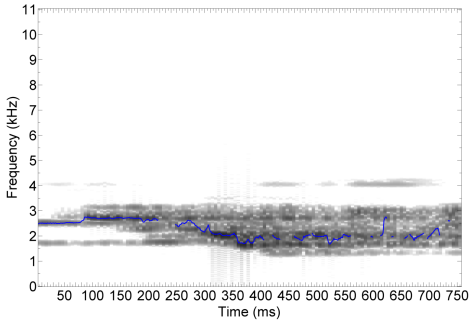


(h) YIN-bird

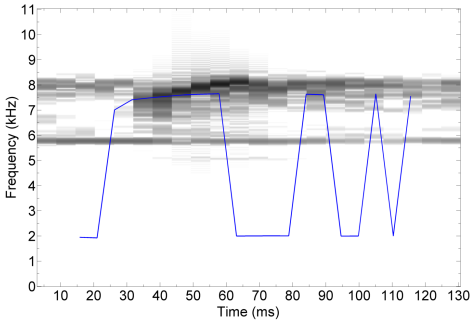
Figure 3.7: Examples of octave errors with YIN. Figures on the left hand side show YIN F_0 contours, while plots on the right hand side show F_0 estimation using YIN-bird with the same examples. F_0 estimates are superimposed in blue.



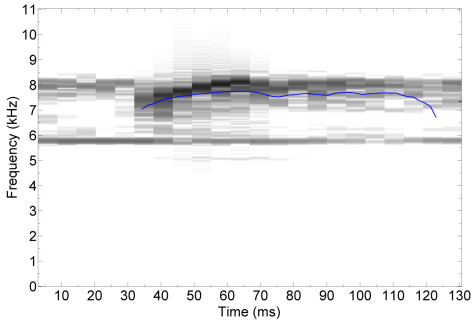
(a) YIN - 02_Yin_buoriole_08_0001



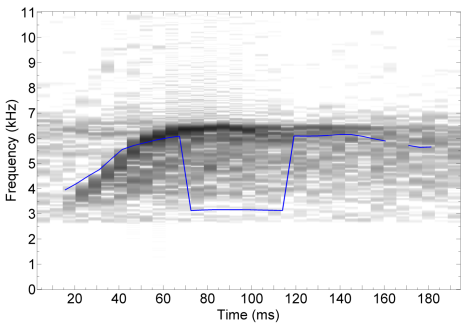
(b) YIN-bird



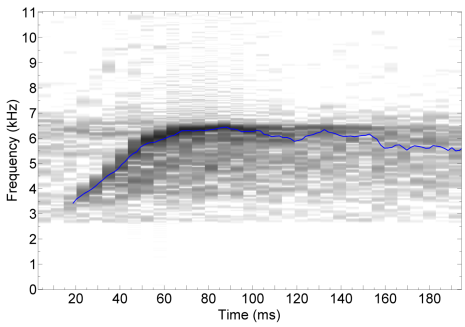
(c) YIN - 12_Yin_buobsb_09_0020



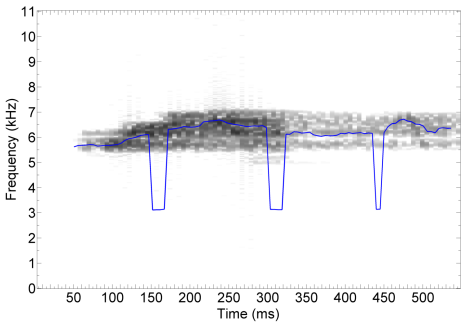
(d) YIN-bird



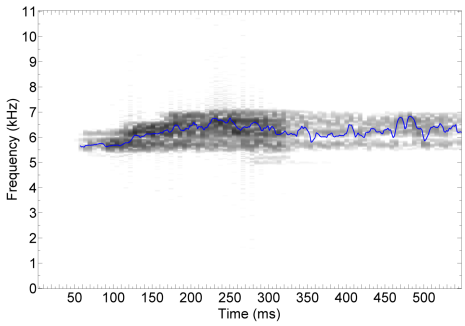
(e) YIN - 11_Yin_buobsb_03_0006



(f) YIN-bird

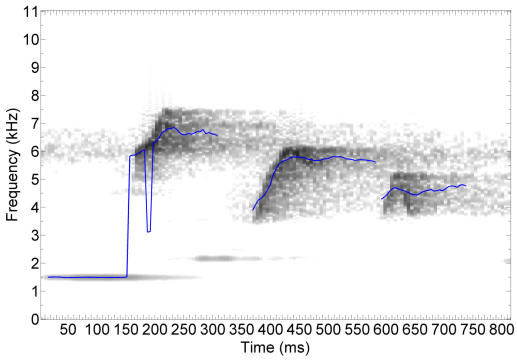


(g) YIN - 18_Yin_wwobsb_03_0036

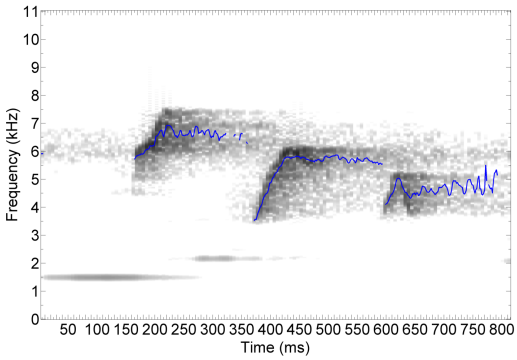


(h) YIN-bird

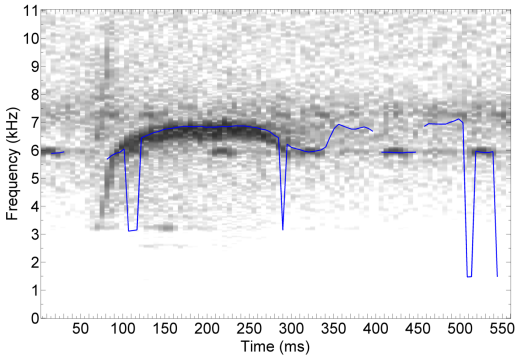
Figure 3.8: Examples of low or degraded quality and short signals. Figures on the left hand side show YIN F_0 contours, while plots on the right hand side show F_0 estimation using YIN-bird with the same examples. F_0 estimates are superimposed in blue.



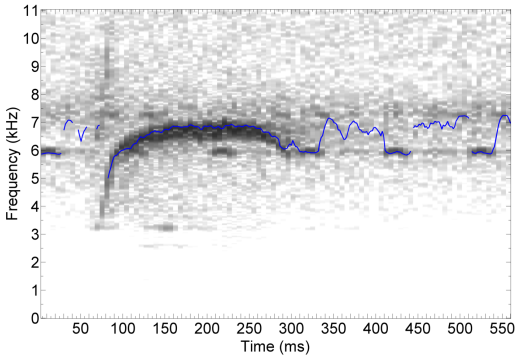
(a) YIN - 09_Yin_buobsb_01_0010



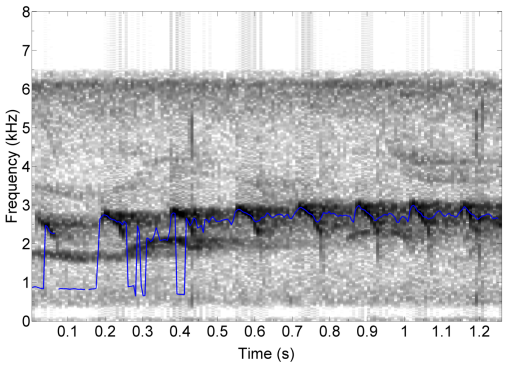
(b) YIN-bird



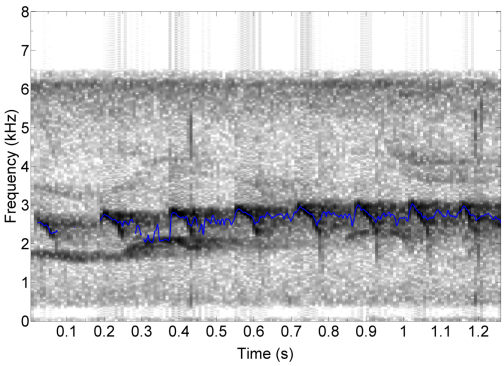
(c) YIN - 14_Yin_suobsb_04_0009



(d) YIN-bird

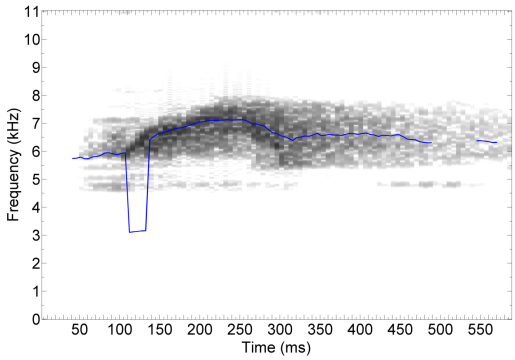


(e) YIN - 22_Yin_D060905K152015

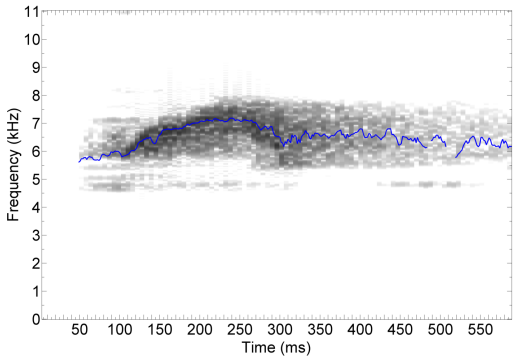


(f) YIN-bird

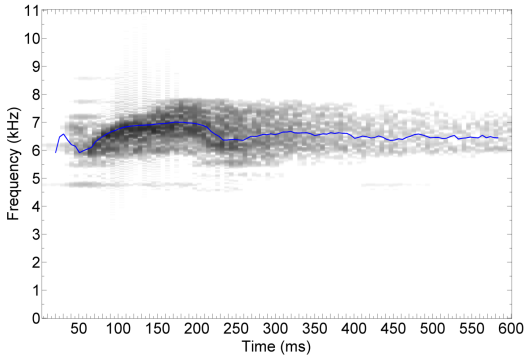
Figure 3.9: Examples of non-target corruption. Figures on the left hand side show YIN F_0 contours, while plots on the right hand side show F_0 estimation using YIN-bird with the same examples. F_0 estimates are superimposed in blue.



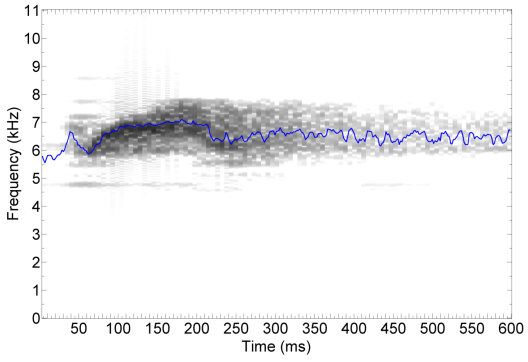
(a) YIN - 15_Yin_wwobsb_01_0001



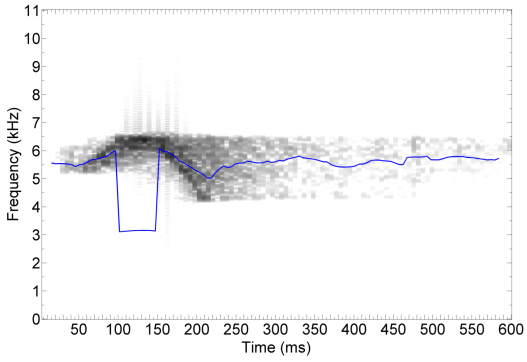
(b)



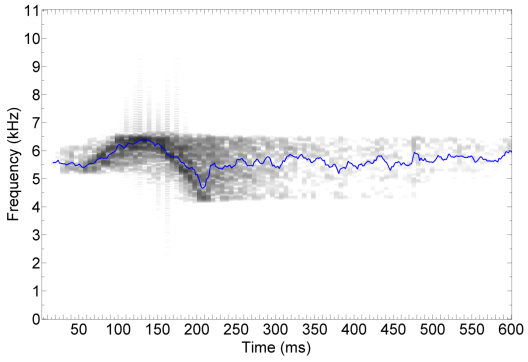
(c) YIN - 16_Yin_wwobsb_01_0003



(d)

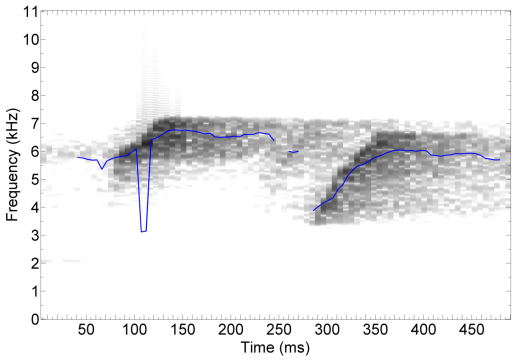


(e) YIN - 17_Yin_wwobsb_02_0003

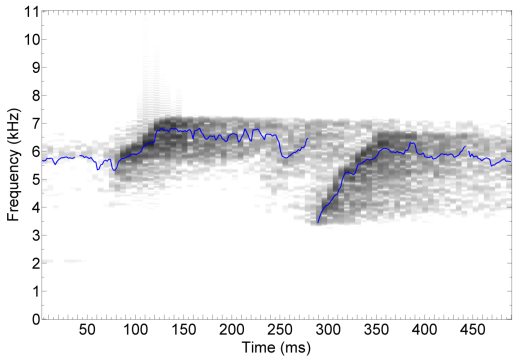


(f)

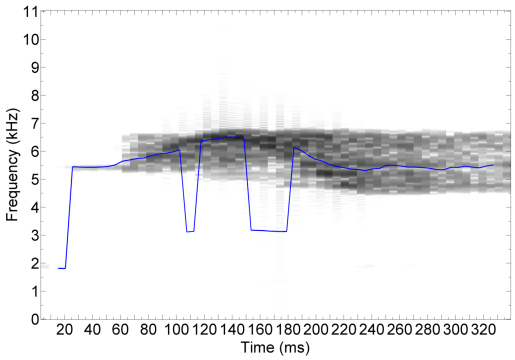
Figure 3.10: Reverb tail examples. Figures on the left hand side show YIN F_0 contours, while plots on the right hand side show F_0 estimation using YIN-bird with the same examples. F_0 estimates are superimosped in blue.



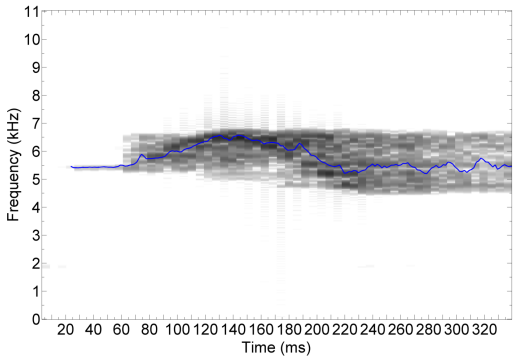
(a) YIN - 07_Yin_buobsb_01_0007



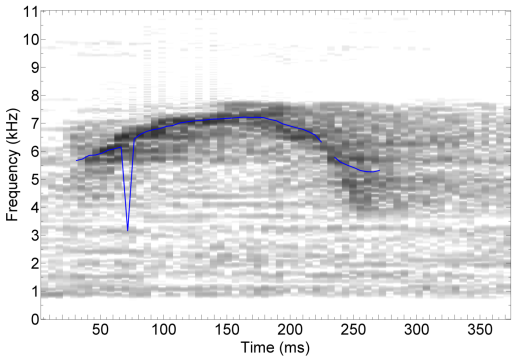
(b)



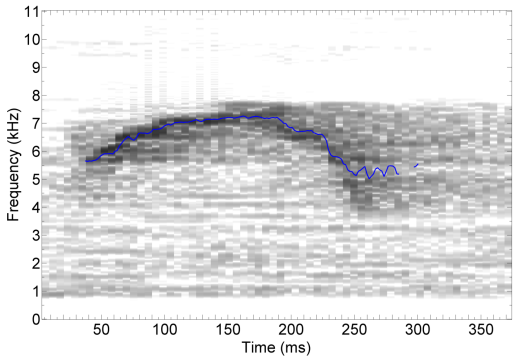
(c) YIN - 20_Yin_wwobsb_07_0014



(d)



(e) YIN - 19_Yin_wwobsb_05_0001



(f)

Figure 3.11: Examples of cleaned up signals subject to octave errors while example of noisier signals (WB noise) is not more prone to octave error. Figures on the left hand side show YIN F_0 contours, while plots on the right hand side show F_0 estimation using YIN-bird with the same examples. F_0 estimates are superimosped in blue.

Table 3.6: Thresholds for bird call difference system with YIN-bird

Set	Data	Threshold 1 (Hz)	Threshold 2 (Hz)
1	BNO	27	60
2	OBSB	71	143
3	AF/BS	93	204
4	antbirds	27	57

3.6 Bird call difference with YIN-bird

The results presented in [116] were described in Section 3.3.1 and Section 3.3.2. Upon development of YIN-bird, a pitch tracking system with improved performance (discussed later in chapter 4), a review of difference measures using YIN-bird was undertaken. These experiments calculated difference scores using this improved pitch tracker. As YIN-bird tracks bird calls pitch more accurately than YIN, a difference measure using YIN-bird should better represent bird population difference.

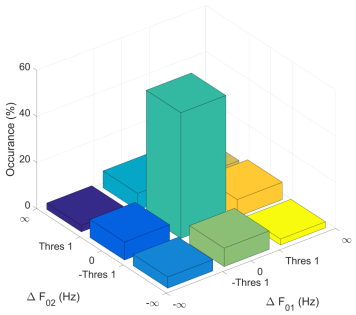
3.6.1 YIN-bird setup for bird difference measure

The first stage, like experiments using YIN, was to extract the $\boldsymbol{F_0}$ contours of the bird data. YIN-bird used the same settings as in [117]. Window size was 6.7 ms, frame rate was 1.7 ms (approximately 75% overlap), quality was set to 'good' which means estimates with aperiodic value of less than 0.2 were considered voiced and segment buffer size was set to 3000 samples (the significance of these parameters are discussed in chapter 4). No minimum $\boldsymbol{F_0}$ parameter was required for YIN-bird as it was adaptively updated to match the spectrogram of the recording.

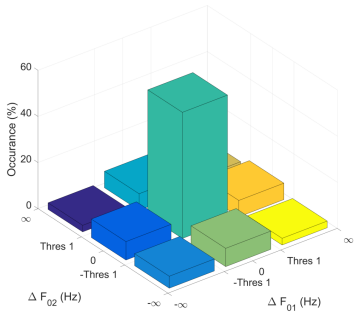
Apart from replacing YIN with YIN-bird the system here was the same as that described in Section 3.3 and displayed in Figure 3.5. With different pitch contour estimates, new thresholds were calculated with the same criteria described in Section 3.3 and Section 3.3.2. The updated thresholds using YIN-bird are shown in Table 3.6.

3.6.2 3x3 codebook difference measure with YIN-bird

Pitch contours were extracted from bird call examples, with $[(F_{02} - F_{01}), (F_{03} - F_{02})]$ or $[\Delta F_{01}, \Delta F_{02}]$ for each population assigned a code based on the 3x3 codebook for the species under test. The number of occurrences per code for each population of the BNO are plotted in histograms of Figure 3.12. Difference scores depend on transition probabilities between codes and not solely on code occurrences, nonetheless the similarity between the histogram in Figure 3.12a and Figure 3.12b indicate the populations are similar. In contrast the histograms of MAT and DWA are quite different as demonstrated in Figure 3.13. These different species should

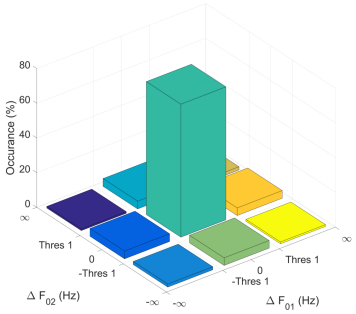


(a) BU BNO - 3x3 code histogram

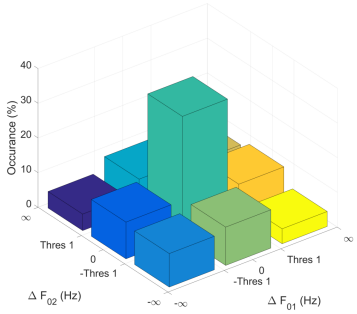


(b) WW BNO - 3x3 code histogram

Figure 3.12: Histograms of 3x3 codebook codes for BNO populations. (a) BU BNO (b) WW BNO. Height of each bar represents the % occurrence of that code. ΔF_{01} divisions are quantized along the x-axis, and ΔF_{02} along the y-axis.



(a) MAT - 3x3 code histogram



(b) DWA - 3x3 code histogram

Figure 3.13: Histograms of 3x3 codebook codes for MAT and DWA species. (a) MAT (b) DWA. Height of each bar represents the % occurrence of that code. ΔF_{01} divisions are quantized along the x-axis, and ΔF_{02} along the y-axis.

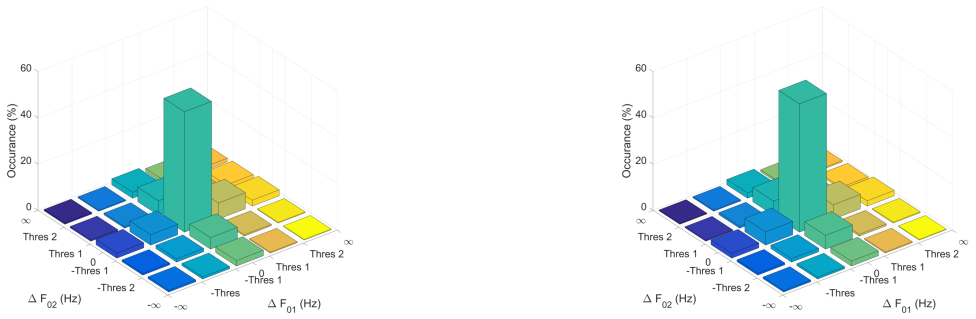
have considerable different vocalizations according to ornithologists.

The difference score results using the same difference measurement system, with YIN-bird replacing YIN, are given in Table 3.7. The change in $\boldsymbol{F_0}$ estimator has dramatically changed the difference scores, but the order is maintained. The smallest score is still BNO with 0.0224. This is expected. AF vs. BS and Mexican antbird scores are consistently higher. The OBSB BUSU vs. WW comparison again falls between these extremes.

Additional comparisons were made during these experiments. The three populations of OBSB were compared to each other. As expected the BU and SU populations score was low as they are the same subspecies, hence little difference. Comparisons between BU & WW, SU & WW and BUSU & WW all output higher difference scores which is in line with zoologist's

Table 3.7: YIN-bird - Bird call difference, 3x3 codebook.

Set	Species	3x3	Description
1	BNO(BU vs. WW)	0.0224	V. similar, different subspecies
2a	OBSB (BUSU vs. WW)	0.1793	WW speciation in dispute, different subspecies
2b	OBSB(BU vs. WW)	0.1736	WW speciation in dispute, different subspecies
2c	OBSB(BU vs. SU)	0.1155	Two populations of same subspecies
2d	OBSB(SU vs. WW)	0.2793	WW speciation in dispute, different subspecies
3	AF vs. BS	0.1569	V. different, different species
4a	DAB vs. DWA	0.1543	Different, same family, different species
4b	MAT vs. DWA	0.8216	V. different, different family, different species
4c	MAT vs. DAB	0.5904	V. different, different family, different species



(a) BU/SU OBSB - 5x5 code histogram

(b) WW OBSB - 5x5 code histogram

Figure 3.14: Histograms of 5x5 codebook codes for OBSB populations. (a) BU/SU (b) WW. Height of each bar represents the % occurrence of that code. ΔF_{01} divisions are quantized along the x-axis, and ΔF_{02} along the y-axis.

thinking the WW OBSB population may have diversified from the other subspecies of OBSB.

Upon evaluation of these experiments the trend appears to be a difference measure of 0 - 0.05 is very similar, 0.051 - 0.13 is similar, 0.131 - 0.2 is different and $>$ 0.21 is very different.

3.6.3 5x5 Codebook difference measure with YIN-bird

Similar to experiments described in Section 3.3.2, YIN-bird pitch contours were used with a 5x5 codebook system. 5x5 code histograms for BU/SU OBSB and WW OBSB are shown in Figure 3.14.

Table 3.8 shows the results. Again, comparisons between Table 3.8 and Table 3.7 are not informative but patterns should still emerge. BNO, as usual, has the lowest difference while MAT vs. DWA has the largest difference. For the OBSB, BUSU vs. WW has a larger difference

Table 3.8: YIN-bird - Bird call difference, 5x5 codebook.

Set	Species	5x5	Description
1	BNO(BU vs. WW)	0.3115	V. similar, different subspecies
2a	OBSB (BUSU vs. WW)	0.986	WW speciation in dispute, different subspecies
2b	OBSB(BU vs. WW)	0.9868	WW speciation in dispute, different subspecies
2c	OBSB(BU vs. SU)	0.9139	Two populations of same subspecies
2d	OBSB(SU vs. WW)	1.6245	WW speciation in dispute, different subspecies
3	AF vs. BS	1.1276	V. different, different species
4a	DAB vs. DWA	0.8494	Different, same family, different species
4b	MAT vs. DWA	2.1122	V. different, different family, different species
4c	MAT vs. DAB	1.4779	V. different, different family, different species

than BU vs. WW again here.

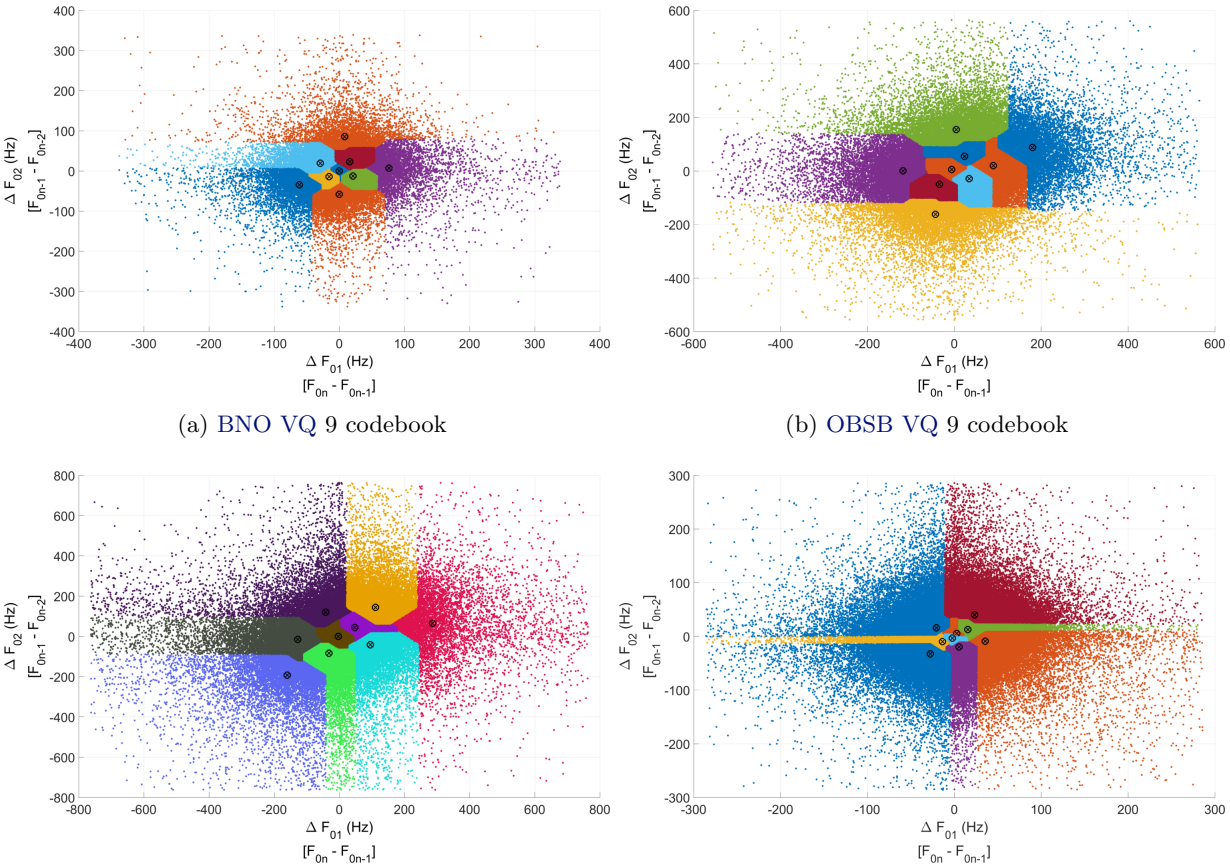
The trend for 5x5 codebook appears to be a difference measure of 0 - 0.4 is very similar, 0.41 - 0.95 is similar, 0.951 - 1.5 is different and $>$ 1.51 is very different. The measure of 0.8494 for DAB vs. DWA is surprising as the expected measure should fall within the bounds of different. Nonetheless the general trend is in line with the difference described by current taxonomy and Zoologists' informed opinions.

3.7 Vector Quantization (VQ) Codebook

The work presented so far in this chapter has dealt with developing a measure of call similarity for birds to aid ornithologists in their work. Continuing on this work, a revised codebook development was implemented using Vector Quantization (VQ). The idea is that instead of uniform 3x3 or 5x5 spacing with a fixed threshold, that for a given species comparison, the centers' locations are chosen based on a VQ process (using k-means). Pitch contour values are then quantized to those centers instead of being assigned a code number from the codebooks described previously in Section 3.3.1 and Section 3.3.2. This allows a more shrewd representation of the space. The remainder of the process remains the same.

3.7.1 VQ codebook per pair - 9 Centers

Using the data from Table 3.1, four VQ codebooks were calculated with 9 centers or clusters. These consisted of one BNO, one OBSB, one AF/BS and one antbird codebook with 9 clusters each. BNO VQ codebook with 9 clusters can be seen in Figure 3.15a. OBSB VQ codebook with 9 clusters can be seen in Figure 3.15b. AF/BS VQ codebook with 9 clusters can be seen in Figure 3.15c. And finally the VQ codebook with 9 centers for the Mexican antbirds can be seen in Figure 3.15d.



(a) BNO VQ 9 codebook

(b) OBSB VQ 9 codebook

(c) AFBS VQ 9 codebook

(d) Mexican antbirds VQ 9 codebook

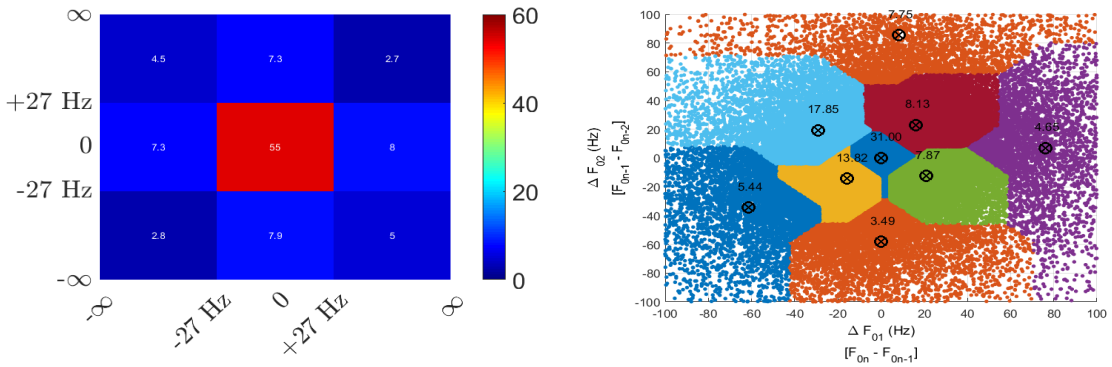
Figure 3.15: Pitch contour VQ 9 codebooks for datasets: (a) BNO (b) OBSB (c) AF/BS (d) Mexican antbirds. ΔF_{01} is plotted on the x-axis, and ΔF_{02} along the y-axis.

Difference measures with these VQ codebooks were calculated using the same process as described in Section 3.6.1. YIN-bird was again used due to its superior performance to YIN (as shown later in chapter 4). Resulting difference measures using this system, where the linear codebooks were replaced by VQ 9 cluster codebooks, are given in Table 3.9. The trend for VQ 9 codebook appears to be a difference measure of 0 - 0.01 is very similar, 0.011 - 0.05 is different and $>$ 0.051 is very different.

The advantage of VQ over a linear spaced codebook is it allows a better distribution of code occurrences. Figure 3.16 shows a comparisons of the distribution of code occurrences in % for the BNO populations. Figure 3.16a shows that the centre code contains 55% of pitch codes whereas in Figure 3.16b, the VQ representation, contours between $\pm 27Hz$ are distributed across 5 clusters.

Table 3.9: YIN-bird - Species specific VQ 9 Codebook

Set	Species	VQ 9	Description
1	BNO(BU vs. WW)	0.0070	V. similar, different subspecies
2a	OBSB (BUSU vs. WW)	0.0352	WW speciation in dispute, different subspecies
2b	OBSB(BU vs. WW)	0.0300	WW speciation in dispute, different subspecies
2c	OBSB(BU vs. SU)	0.0053	Two populations of same subspecies
2d	OBSB(SU vs. WW)	0.0589	WW speciation in dispute, different subspecies
3	AF vs. BS	0.0870	V. different, different species
4a	DAB vs. DWA	0.0119	Different, same family, different species
4b	MAT vs. DWA	0.121	V. different, different family, different species
4c	MAT vs. DAB	0.0678	V. different, different family, different species



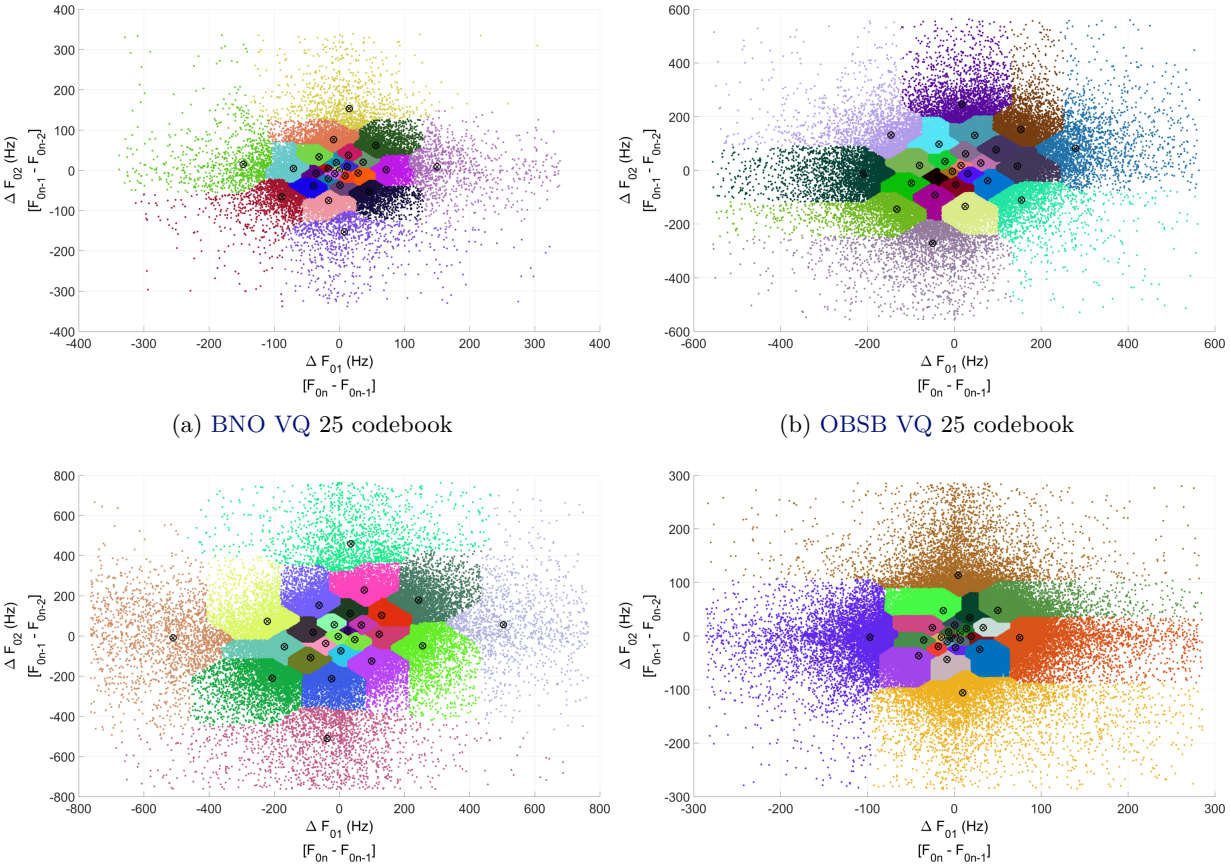
(a) BNO 3x3 distribution

(b) BNO VQ distribution

Figure 3.16

3.7.2 VQ codebook per pair - 25 Centers

A more detailed representation of the pitch slopes was investigated when the 3x3 codebook was extended to a 5x5 codebook in Section 3.3.2. A similar approach was used here with VQ. The number of clusters was increased from 9 to 25, to allow a more detailed representation of the pitch contour. Again four codebooks were constructed, one BNO, one OBSB, one AF/BS and one antbird codebook, but with 25 centers this time. These codebooks can be seen in Figure 3.17. Resulting difference measures using this expanded VQ 25 codebook system, are given in Table 3.10. The trend for VQ 25 codebook appears to be a difference measure of 0 - 0.01 is similar, 0.011 - 0.1 is different and $>$ 0.11 is very different.



(a) BNO VQ 25 codebook

(b) OBSB VQ 25 codebook

(c) AFBS VQ 25 codebook

(d) Mexican antbirds VQ 25 codebook

Figure 3.17: Pitch contour VQ 25 codebooks for datasets: (a) BNO (b) OBSB (c) AF/BS (d) Mexican antbirds. ΔF_{01} is plotted on the x-axis, and ΔF_{02} along the y-axis.

Table 3.10: YIN-bird - Species specific VQ 25 Codebook

Set	Species	VQ 25	Description
1	BNO(BU vs. WW)	0.0096	V. similar, different subspecies
2a	OBSB (BUSU vs. WW)	0.0732	WW speciation in dispute, different subspecies
2b	OBSB(BU vs. WW)	0.0746	WW speciation in dispute, different subspecies
2c	OBSB(BU vs. SU)	0.0092	Two populations of same subspecies
2d	OBSB(SU vs. WW)	0.0742	WW speciation in dispute, different subspecies
3	AF vs. BS	0.1188	V. different, different species
4a	DAB vs. DWA	0.0653	Different, same family, different species
4b	MAT vs. DWA	0.2551	V. different, different family, different species
4c	MAT vs. DAB	0.3804	V. different, different family, different species

3.8 Data size and self comparison

3.8.1 3x3 & 5x5 codebook with reduced data size and self comparison

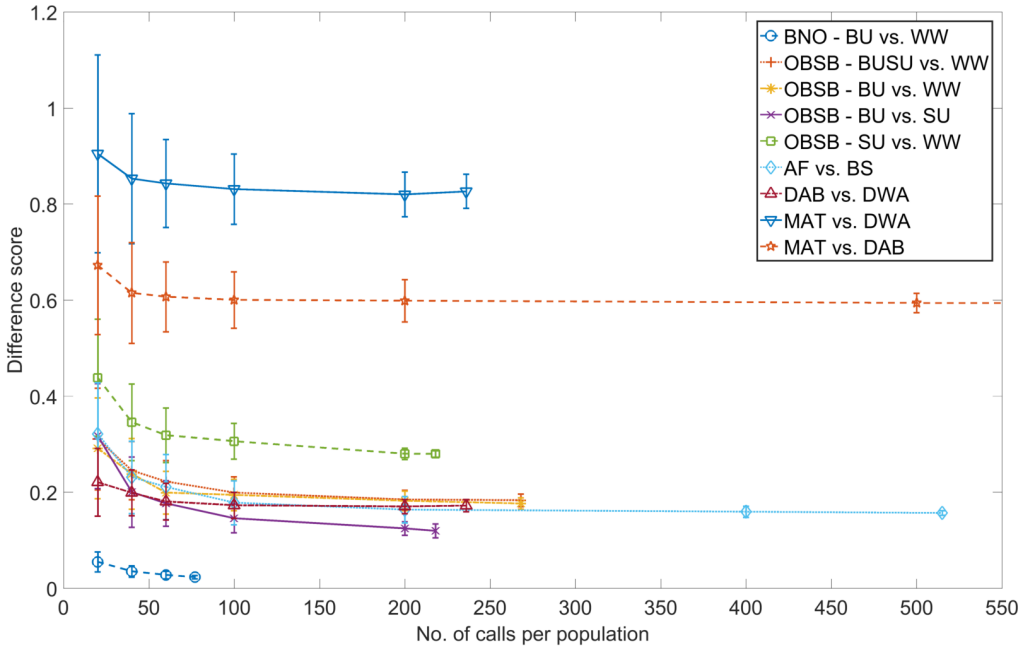
To investigate the influence data size has on the vocalization difference measure, subsets of population data were used to detail it's influence. This was to address the question, "How many examples of a bird population are needed to be confident in a difference measure?". The BU BNO has 110 examples and WW BNO has 77 examples. Previous comparisons used all available data, i.e. 110 samples vs. 77. Here subsets with the same size for both populations were used. Subset sizes for BNOs were 20, 40, 60, and 77. A vector containing random permutation of integers from 1 to population size (n), with subset size (k) unique integers was generated. This was done 100 times per subset size to allow averaging. In the case of BU and WW BNO, each population had 4 subset sizes (20, 40, 60, 77), each with 100 different permutations of k between 1 - n . Using these vectors as a seed, in the case of subset size equal to 20, 20 BU BNO and 20 WW BNO examples were evaluated using the 3x3 difference measure (using YIN-bird) 100 times and the average is plotted in Figure 3.18a with error bars indicating the standard deviation. Note how averages converge and error bars reduce when number of samples increases. Figure 3.18b plots results using the 5x5 codebook. Codebooks generated for previous experiments were used here again for consistency, although the codebooks were calculated using data that might not necessarily be used in a given subset experiment i.e. all available data for a species comparison was used when calculating thresholds, not all available data was used in each subset. The thresholds per species were given in Table 3.6.

If populations which are similar have a low difference score, then subsets of the same population should also have low difference scores. This theory was tested. This time a vector containing random permutation of integers from 1 to population size (n), with n unique integers was generated, i.e. numbers were jumbled up. For a given population, subset A contained the first k examples from the random permutation seed and subset B contained the last k examples from the random seed. This resulted in two subsets of the same populations with the same number of examples but only one group contained any given example. Again this process was done 100 times for different subset sizes. The average difference scores of these self comparisons are plotted in Figure 3.19a using 3x3 codebooks and Figure 3.19b using 5x5 codebooks, with error bars expressing standard deviation.

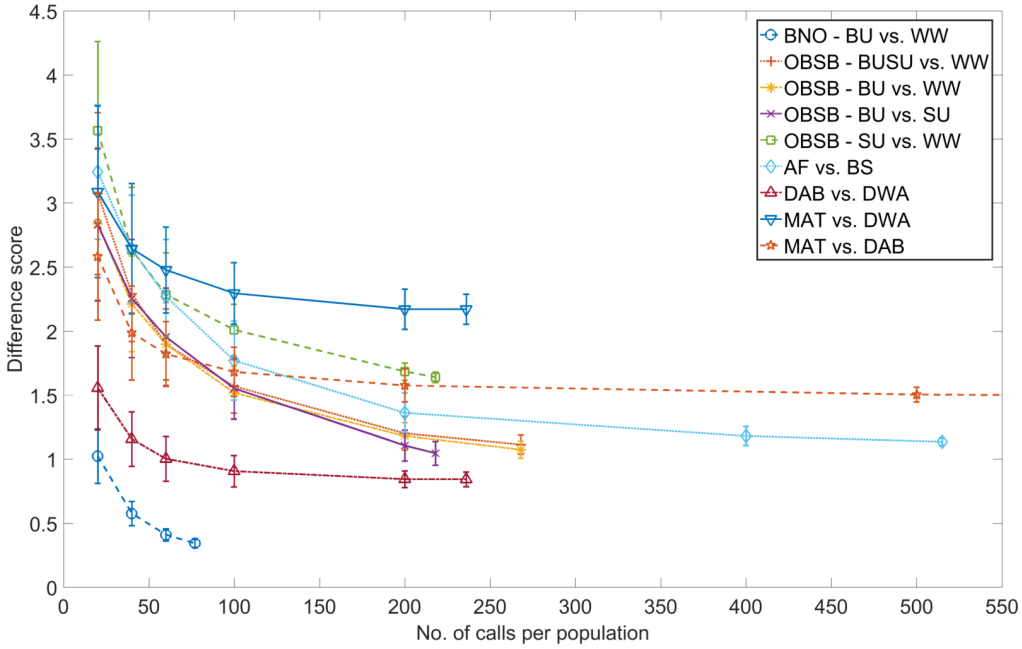
3.8.2 VQ codebook with reduced data size and self comparison

The experiments from Section 3.8.1 were repeated using the VQ codebooks from Figure 3.15 and Figure 3.17. The results of reduced subset comparisons using VQ codebook with 9 centers are potted in Figure 3.20a and with 25 centers in Figure 3.20b.

The self comparison results of populations using the bird vocalization difference measure are shown in Figure 3.21a employing the VQ codebooks with 9 centers and Figure 3.21b using the

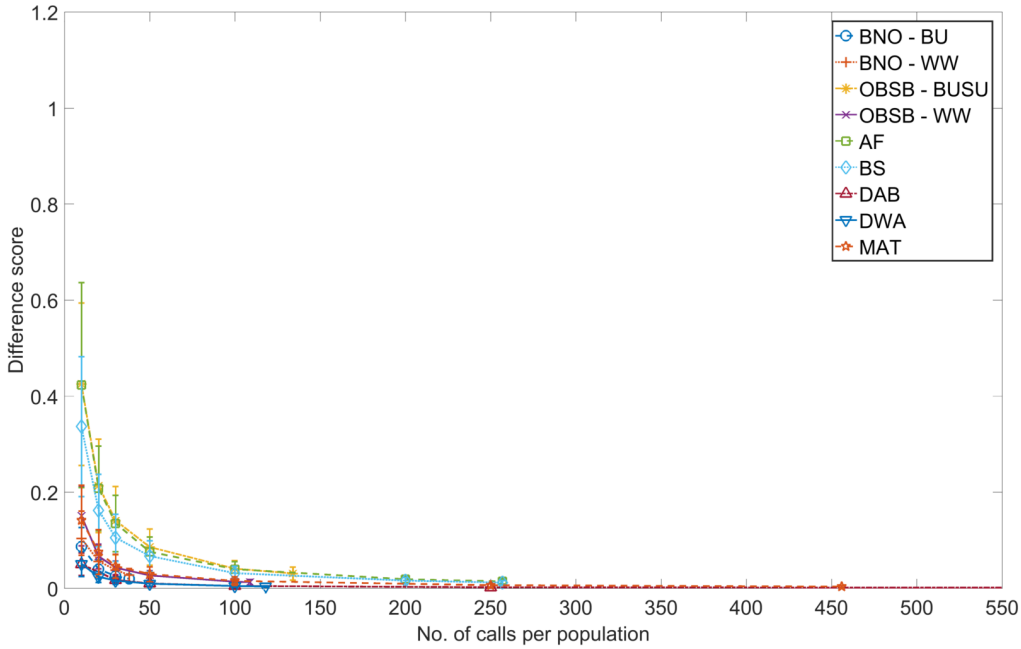


(a) Comparisons using reduced data sets - 3x3 codebook

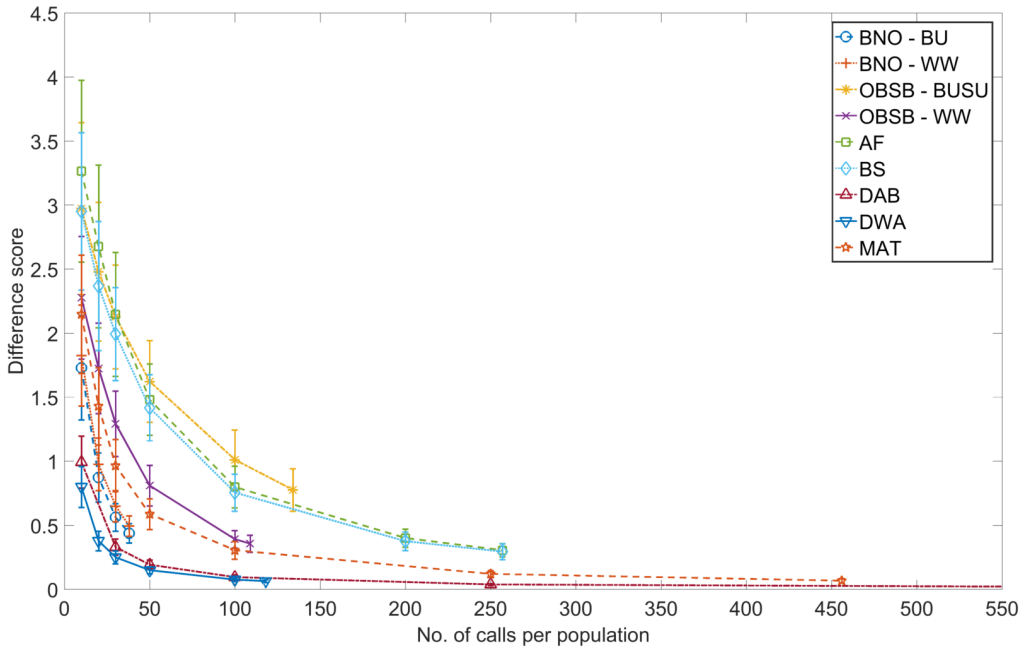


(b) Comparisons using reduced data sets - 5x5 codebook

Figure 3.18: Comparisons using reduced data per population with 3x3 and 5x5 codebooks. The number of bird call examples per population is plotted on the x-axis with the difference score on the y-axis.



(a) Comparisons using subsets of the same population - 3x3 codebook



(b) Comparisons using subsets of the same population - 5x5 codebook

Figure 3.19: Comparisons using two subsets of the same population with 3x3 and 5x5 codebooks. These scores show low difference measures between different subsets of the same population. The number of bird call examples per population is given on the x-axis with the difference score on the y-axis.

25 center VQ codebooks. Again YIN-bird was used to extract pitch for these examples. These plots show comparisons using subsets of the same population with mean and variance of the 100 trials per subset size given by a marker with error bars. These self comparisons indicate more data may be required when using the VQ process as error bars are quite large when data size is small. AF and BS self comparisons are slightly higher than expected but still a lot lower than the difference score for AF vs. BS (0.087). AF self comparison for 257 calls is 0.01156 and for BS is 0.006193 with VQ 9, which is similar to DAB vs. DWA score of 0.0119, but still much smaller than 0.121 for MAT vs. DWA. Self comparison score for MAT with 250 calls is 0.001104 a factor of 100 more similar than MAT vs DWA (DWA with 118 calls self score is 0.001935).

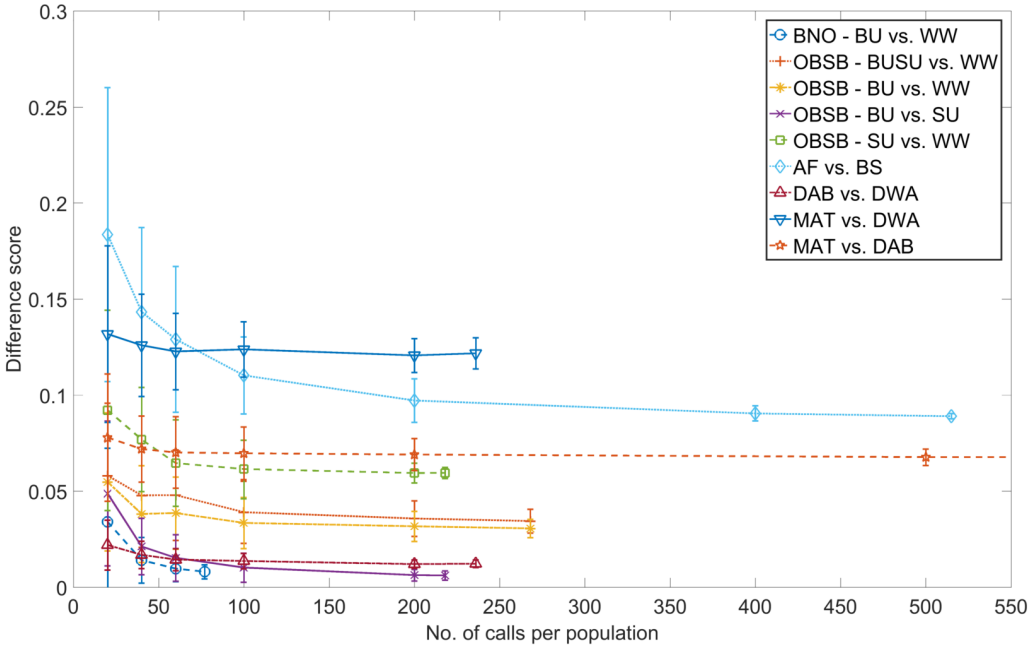
Trends are similar for self comparisons using VQ 25. AF and BS self comparisons are again slightly higher than expected but still a lot lower than the difference score for AF vs. BS (0.119). AF self comparison for 257 calls is 0.051 and for BS is 0.043 with VQ 25, which isn't too much smaller than DAB vs. DWA score of 0.065, but still much smaller than 0.2551 for MAT vs. DWA. Self comparison score for MAT with 250 calls is 0.014 a factor of 18 more similar than MAT vs DWA (DWA with 118 calls self score is 0.016).

3.9 Discussion

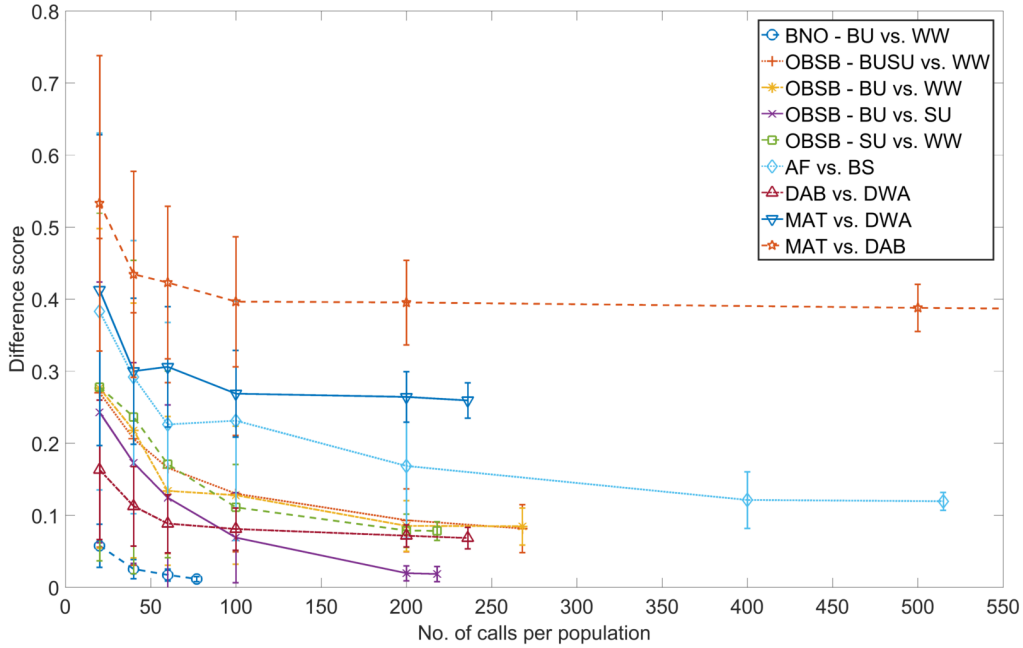
Ideally, ornithologists would like a measure and methodology to compare calls or songs from different populations that allows them to say that the vocalizations are very similar or quite different. It may not be possible to exactly pinpoint where the threshold between similar and different lies. Such evidence is always taken in parallel with genetic and morphological data about the bird population to build a complete picture. Work here aimed to identify regions of confidence for similarity and difference with a boundary region in between. This requires more development of the current approach. These initial results are promising however. The measure for the overlapped BNO populations is smaller than it is for the antbird populations which are known to be diverse and from separate species.

The difference scores using YIN will not be discussed, as the experiments using YIN-bird are intended to be a replacement of initial experiments using a more accurate pitch tracking system for birds. Using 3x3 codebooks, the BNO difference measure is the lowest for all subset sizes. The comparisons between BU, BUSU, SU OBSB and WW OBSB each have a higher difference score than the comparison between BU and SU OBSB. This is expected as the BU and SU populations are considered the same species and therefore should have similar vocalizations. The WW OBSB is currently a different subspecies but following ongoing analysis by zoologists may be reclassified as a new separate species. AF vs. BS score was surprisingly lower than expected. It was still higher than subspecies comparisons but not by much at 0.1569, with OBSB BU vs. SU being 0.1155.

Experiments using reduced datasets per population show that the mean difference measures converge at about 100 samples per population with medium variance (see Figure 3.18a). Per-

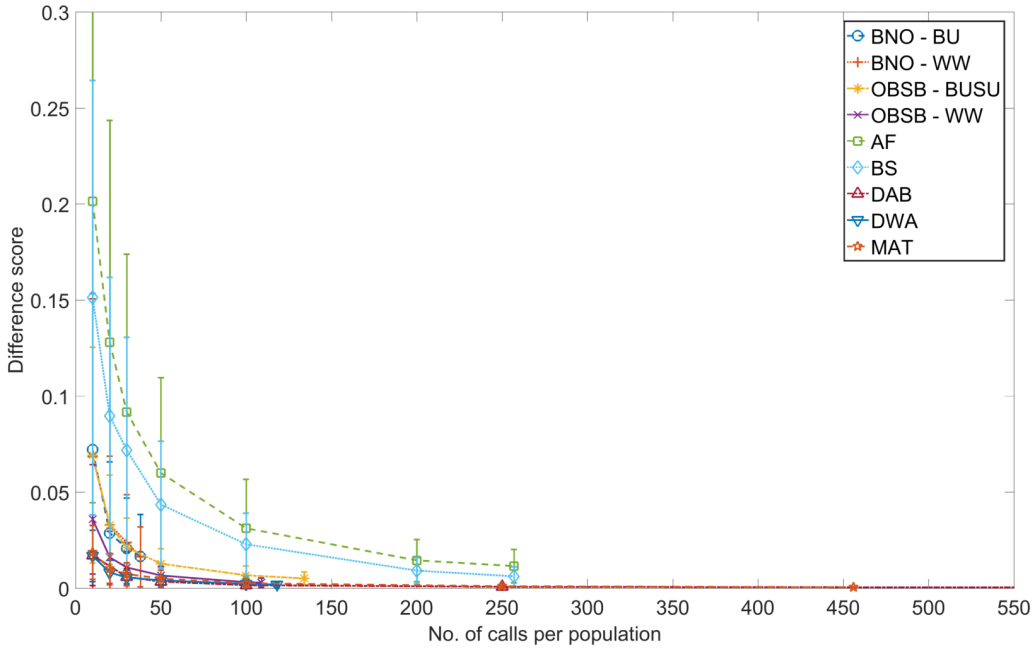


(a) Comparisons using reduced data sets - VQ 9 codebook

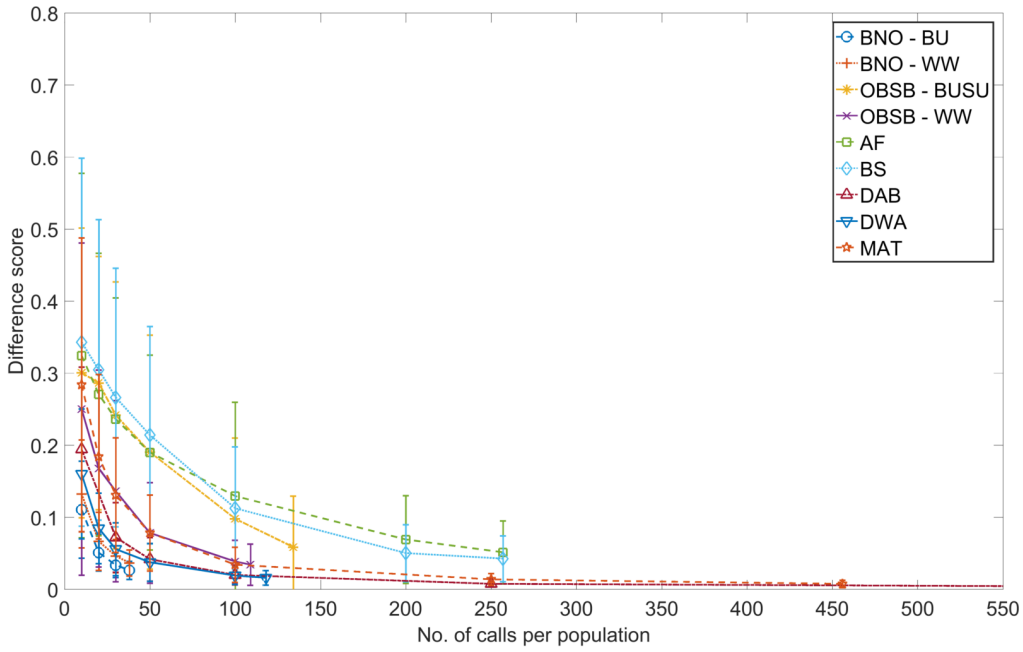


(b) Comparisons using reduced data sets - VQ 25 codebook

Figure 3.20: Comparisons using reduced data per population with VQ codebooks. The number of bird call examples per population is plotted on the x-axis with the difference score plotted on the y-axis.



(a) Comparisons using subsets of the same population - VQ 9 codebook



(b) Comparisons using subsets of the same population - VQ 25 codebook

Figure 3.21: Comparisons using two subsets of the same population with VQ codebooks. These scores show low difference measures between different subsets of the same population. The number of bird call examples per population is plotted on the x-axis with the difference score on the y-axis.

forming difference comparisons with different subsets of the same populations show difference measures comparable with the BNO score which expresses them to be very similar as expected (see Figure 3.19a).

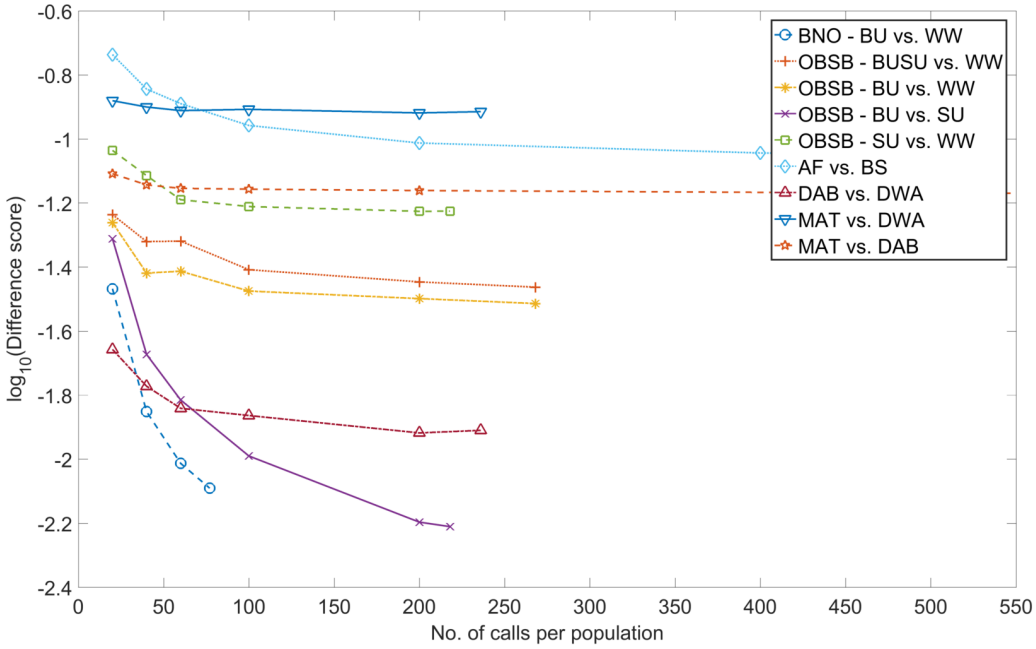
Moving to the 5x5 codebook made the difference between the AF and BS species larger, suggesting greater diversity in their calls. However, the data in this set was the lowest quality data used and it is likely the BS data contained multiple subspecies as it was not possible to get enough data from a single geographical location. Control over and access to data is an ongoing challenge in this development. Intuitively the use of the 5x5 codebook allows capture of a more complex micro-structure of pitch contour. Whether this is necessary for all bird population comparisons is not clear. For instance, the BNO's call is the simplest of the calls used here hence the pitch contour varies less. A 3x3 codebook may be sufficient for the BNO whilst not for others. It is difficult to quantify an improvement from the 3x3 to 5x5 comparisons as no ground truth difference exists. Both metrics align with ornithologists qualitative opinions of difference. Perhaps using difference regions instead of continuous numbers for the final mark (similar to the work of Tobias et al. [160]), is a good solution. Although this would also require suitably labelled data from ornithologists to benchmark against. Figure 3.18b shows mean and variance of difference scores using reduced data with 5x5 codebooks. Difference scores start to converge around 200 samples per population implying the addition of a second threshold requires more data when comparing populations. Figure 3.19b shows difference scores between subsets of same populations. Error bars show within species variance for different subset sizes. Again at approximately 200 calls per subset the difference score converges and these scores are within the region of very similar, just like the score for the BNO using a 5x5 codebook.

A better representation of the pitch contour space using Vector Quantization was then implemented. This replaces the uniformly spaced codebook with a more flexible representation that can be tuned for a given species comparison. The thresholds for 3x3 and 5x5 codebooks are quite high and may not allow a sufficiently nuanced division of the pitch microstructure space. A VQ codebook allows centers to be placed at important regions within the microstructure space. This should lead to more accurate difference comparisons. The results in Table 3.9 and Table 3.10 appear to be more in line with species difference described by zoologists when compared to results using 3x3 and 5x5 codebooks. Figure 3.20a shows difference measures with VQ codebooks of 9 clusters for various data sizes. The difference measure converges between 100-200 calls per population with low variance shown by tight error bars. Figure 3.20b shows difference measures with VQ codebooks of 25 clusters for various data sizes. The difference measure for this codebook converges between 100-200 calls per population also. Using VQ with 9 centers, vocal difference scores below 0.01 resemble similar subspecies and above 0.01 resemble different species. For VQ with 25 centers vocal difference scores below 0.04 resemble similar subspecies and above 0.04 resemble different species. Difference scores using VQ 25 codebooks also converge between 100 - 200 calls per population. Difference scores using subsets of the same population with VQ 9 are shown in Figure 3.21a. Most populations' difference scores converge

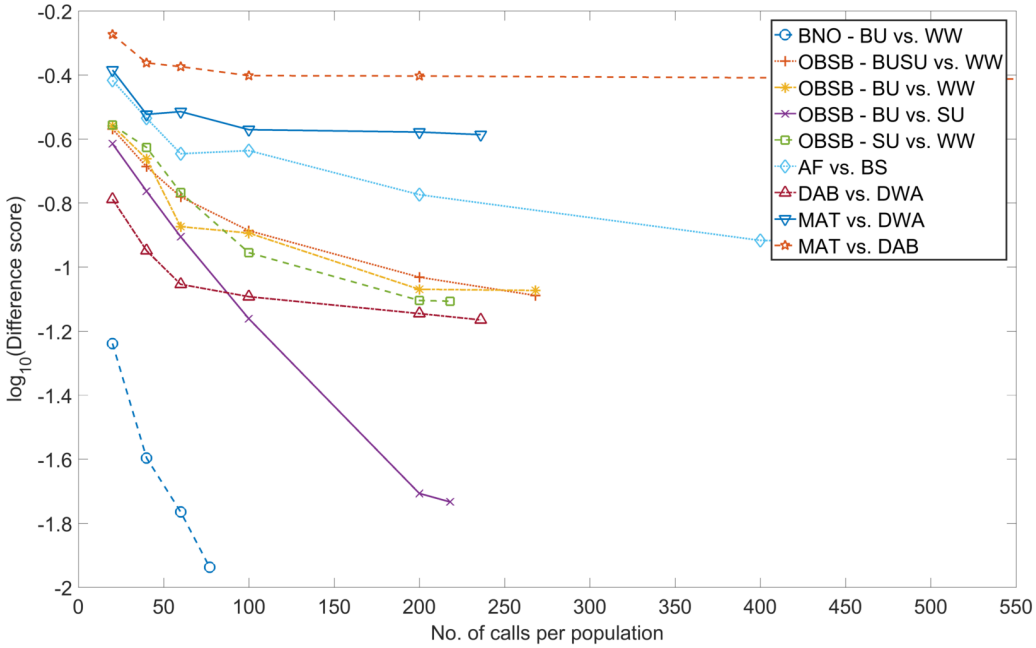
at 100 samples per subset with a difference score less than 0.01, but AF and BS self comparisons take more data to converge and within population difference score is slightly higher than 0.01 with 200 samples per subset. The same is true when using VQ codebooks with 25 centers for self comparisons (see Figure 3.20b), AF and BS self comparisons take more data to converge but most populations' difference scores converge around 200 calls per subset and are less than 0.04. Ornithologists need to know how many recordings they need to take in the field to later analyze bird vocalizations as such data is not easy to collect. Knowing convergence of difference scores helps propose guidelines as to how many calls are necessary.

Finally the difference measures computed using VQ 9 and VQ 25 codebooks are plotted on a log_{10} scale in Figure 3.22a and Figure 3.22b for clarity. Using VQ 9, BNO and BU vs. SU OBSB comparisons have the lowest difference score with BU vs. WW OBSB, BUSU vs. WW OBSB, SU vs. WW OBSB, MAT vs. DAB, AF vs. BS, and MAT vs. DWA having higher levels of difference. DAB vs. DWA score is somewhere in between. A similar trend is seen in Figure 3.22b using VQ 25 codebooks. These results are promising as they are consistent with levels of difference predicted by zoological taxonomies.

This chapter has presented a measure of similarity for bird calls and songs based on the difference in distribution of pitch contours. Whilst still undergoing development, the reported experiments give confidence that a system can be developed that quantifies difference of bird vocalizations. With more research, a difference threshold region can be determined to aid ornithologists in the species/subspecies decision making process.



(a) Comparisons using reduced data sets - VQ 9 codebook



(b) Comparisons using reduced data sets - VQ 25 codebook

Figure 3.22: Log scale comparisons using reduced data per population with VQ codebooks. The number of bird call examples per population is plotted along the x-axis with the difference score plotted on the y-axis.

4

YIN-bird: Improved pitch tracking for birds

Pitch or Fundamental Frequency (F_0) is an important feature of bird song, from which scientists can learn much about a population. To use pitch as a feature, researchers need confidence in their pitch extraction system. Pitch Detection Algorithms (PDAs) proven to work on human speech may not be suitable for all types of bird vocalizations. Chapter 4 discusses pitch estimation performance on a variety of common bird vocalizations. The presence of multiple partials or tones simultaneously, extended frequency sweeps through multiple octaves, and rapid pitch modulations are just some of the difficulties encountered when estimating the pitch of bird song. Carefully tuned parameters improve pitch tracking with YIN, but optimal parameters can change quickly even within one song. This chapter also presents YIN-bird, a modified version of popular pitch detection algorithm YIN which exploits spectrogram properties to automatically set a minimum F_0 parameter for YIN. Gross pitch errors on whistles and trills see a relative improvement by up to 65% on a ground truth dataset of synthetic bird song with known pitch. This dataset was evaluated by expert listeners and described as "sounding like original & can hardly tell it is synthetic". A qualitative analysis of complex bird vocalizations is also presented.

4.1 Pitch Tracking

Again an important theme of this thesis is determining how different two populations of birds are based on their calls and songs. In chapter 3 (and [116] by O'Reilly et al.), pitch contour micro-structure was used to measure similarity of bird calls and songs inspired by dialect similarity measures used in [102] & [103]. This system was reliant on pitch tracking accuracy.

Because of the importance of pitch to human perception, there have been countless studies on automatic methods of extracting F_0 and a brief review is given in this Section. Pitch tracking of bird song can be considered quite specialist within the bird community and improving the ease of which pitch analysis can be done would be beneficial. YIN [42], has been successfully used to detect pitch in human speech. Preliminary experiments for work in this chapter showed that YIN has strong potential for pitch tracking in bird song too. However, it must be carefully tuned for each species and often even for different segments of a single song. This chapter presents a modification to YIN to allow more fully automated pitch tracking. This offers advantages in large batch processing where outputs can't be checked in detail. The aim is to offer zoologists a tool for pitch tracking that requires less specialist knowledge and intervention. Improving automatic pitch estimation of bird vocalizations is also beneficial to engineers and scientists, allowing larger scale studies where results can achieve greater levels of statistical significance. Knowledge of performance of current pitch tracking systems is also important. In many previous experiments, such as [101] and [137], researchers may have used pitch tracking systems designed for human vocalizations and assumed the accuracy to be sufficient.

4.1.1 Pitch Importance

Pitch or F_0 estimation is a much debated topic in speech processing. In speech, the term Fundamental Frequency describes the period of voiced speech. Pitch is the perception of F_0. F_0 is the inverse of the smallest true period in the interval being analyzed [152]. Pitch is the perceived F_0 of a signal [23]. A sound which may not be periodic can still arouse a pitch, but over a wide range, period and pitch are considered equal and F_0 estimation methods are often referred to as Pitch Detection Algorithms (PDAs) [42].

As mentioned in Section 2.8, pitch is very useful for bird song analysis. Pitch provides important information about a sound source. In speech, pitch can be used for a variety of tasks like identifying gender, as males tend to speak with a lower F_0 than women [170]. Speaker emotion can be inferred from pitch, e.g. low pitch can suggest the speaker is sad while high pitch suggests excitement. Pitch changes in a sentence influence how the sentence is interpreted, e.g. a rising pitch is generally observed when a question is asked [107]. In music, pitch estimation is used to name notes [144] which can be used for automatic music transcription [9]. Pitch is also used to identify the artist and title of recordings from short clips [169].

4.1.2 Pitch Tracking Techniques

Throughout the last 30 years, development of PDAs has been a hot research topic for engineers. While there have been major advancements with many PDAs, debate still surrounds which tool to use under certain conditions. This section gives a brief overview of commonly used algorithms. It is not intended as a comprehensive review of all available algorithms. For in depth comparisons the reader is directed to [152, 42, 144, 23, 92].

Most PDAs use a time domain approach (e.g. autocorrelation) or a spectral approach (e.g. cepstrum or harmonic relations). Which algorithm to use depends on the input signal and the conditions it was recorded in. For instance, when tracking pitch of telephone speech, Harmonic Product Spectrum (HPS) [140] PDA will fail if the F_0 has been filtered out or other harmonics are missing. Subharmonic Summation (SHS) PDA [62], which uses addition instead of multiplication, overcomes the problem of missing harmonics, but subharmonics could be falsely predicted as F_0 when using SHS. Subharmonic-to-harmonic Ratio (SHR) PDA [151] improves upon HPS and SHS. While HPS and SHS only look at harmonics of the F_0, SHR adds harmonics as well as subtracting the spectrum at the middle points between harmonics. Like HPS and SHS, SHR cannot recognize pitch of inharmonic signals.

YIN PDA is based on the well-known autocorrelation method with a number of modifications [42], Autocorrelation (AC) based pitch estimators are preferred to HPS, SHS and SHR for the majority of cases as AC can deal with missing harmonics and inharmonic signals and with some modifications (like those in YIN), AC can overcome the problem of giving high scores for subharmonics of the pitch. YIN uses average squared difference instead of AC and includes several modifications that combine to prevent errors. YIN looks for dips instead of peaks (which is why it's called YIN as opposed to YANG) which makes it immune to amplitude changes which affect AC [42]. A example of F_0 estimation using the AC function and YIN is shown in Figure 4.1 [42].

Finally Cepstrum (CEP) [110] is a pitch extraction method similar to AC but CEP estimates the pitch as the frequency whose inverse maximizes the cepstrum of the signal. CEP appears less frequently in pitch tracking literature, but is included here for completeness.

Most pitch tracking software uses some form of the discussed techniques. *Praat* [15] is a commonly used package that estimates pitch in two ways, autocorrelation and cross-correlation. RAPT (sometimes referred to as GET_F0) [152] is a robust algorithm that uses a multi rate approach and the Normalized Cross-correlation Function (NCCF). eSRPD [6] is a super resolution pitch determinator that uses NCCF and removes discontinuities during post-processing. TEMPO [74] is a tool found in STRAIGHT [73], a speech analysis and synthesis toolkit, and uses the instantaneous frequency of the outputs of a filterbank. It estimates pitch using both time interval and frequency cues, and is designed to minimize perceptual disturbance due to errors in source information extraction [5].

Other pitch trackers worth noting are, Summation of Residual Harmonics (SRH) [47] and SWIPE' [23]. SRH is a method based on the summation of residual harmonics exploiting a spectral criterion on the harmonicity of the residual excitation signal. SWIPE' [23] is a 'Sawtooth Waveform Inspired Pitch Estimator' developed for processing pitch in speech and music, and is shown to outperform existing algorithms on speech/musical-instruments databases and a disordered speech database.

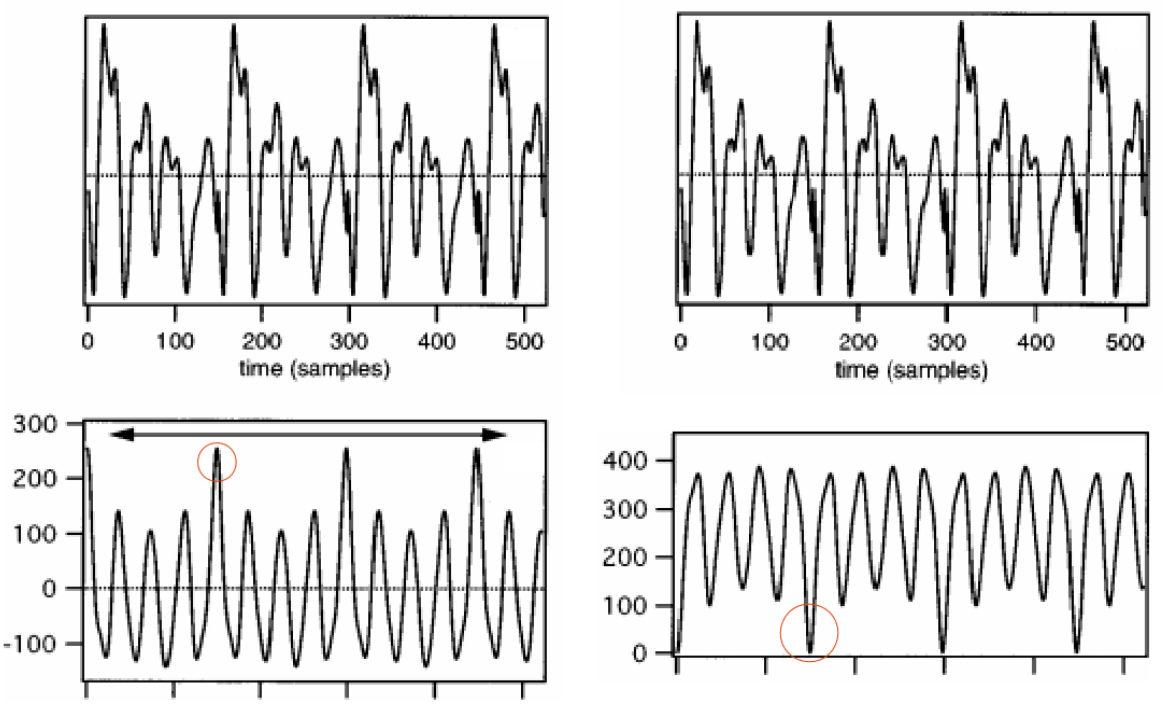


Figure 4.1: Left hand side of this figure shows a sample of speech (top) and its AC (bottom), the right hand side shows squared difference function (which is the algorithm YIN uses) of the same speech sample. An orange circle represents the peak of the AC function and the dip of the difference function, which is used to estimate the pitch [42].

4.1.3 Current pitch tracking performance

De Cheveign and Kawahara in [42] presented an evaluation of YIN against AC, Cross-Correlation, SHS, eSRPD, CEP & TEMPO. YIN outperformed these (and other) pitch estimators when tested on Japanese, English and French databases.

Luengo et al. in [92] showed *Praat*, RAPT and CEP to work better than SHR on clean speech and noisy speech. Work by Wei and Alwan [171] showed *Praat* to outperform RAPT (GET_F0) and TEMPO under both white noise and babble noise conditions (YIN was not included in this study).

Evaluation of pitch tracking performance by Camacho [23] showed SWIPE' to perform the best followed by SHS, RAPT, TEMPO and YIN with gross errors (lower gross error is desired) less than 2.10% when tested on clean speech. SHR, eSRPD & CEP had a gross error greater than 3.5%. Pitch estimation was also trialled on musical instruments in [23]. The gross error rate for tests on musical instruments by octave for SWIPE' was 0.97% and YIN was 0.99%. SHR performed the worst at 36%. RAPT was excluded from musical instrument tests because the bandwidth of musical instruments is too large for the two pass down sampling method used by RAPT.

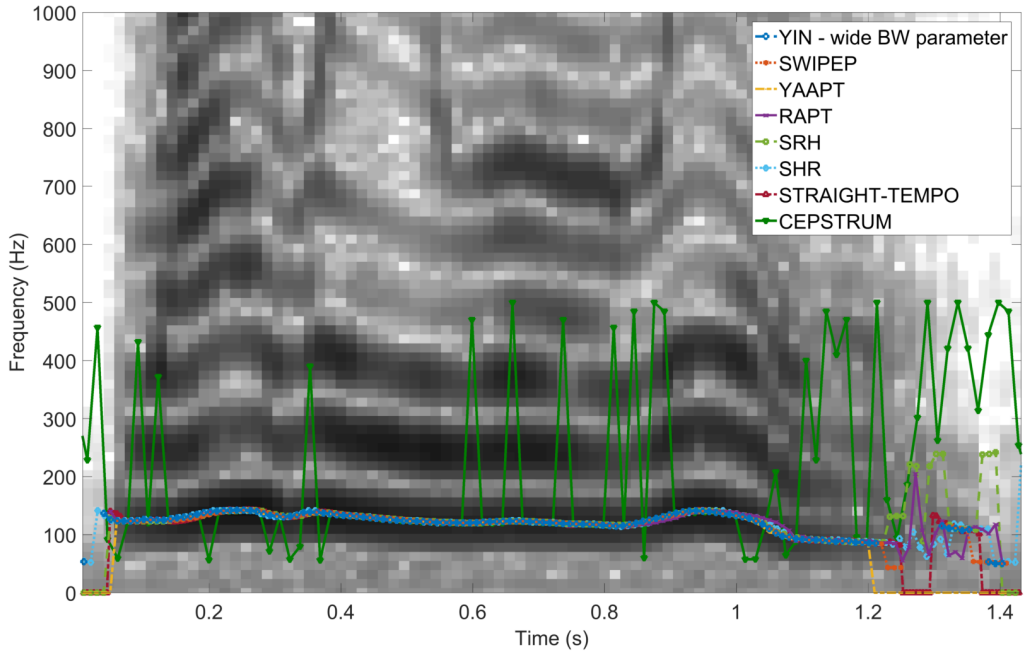


Figure 4.2: Pitch tracking on speech example from TIMIT database.

Babacan et al. [5] evaluated pitch tracking on singing voice with *Praat* and RAPT providing the best determination of voicing boundaries. RAPT reached the lowest number of gross pitch errors while YIN achieved the second best gross pitch error on singing. YIN also had the best fine pitch error on singing sounds. These standard error metrics are described in Section 4.3.4. Finally, YIN was shown to suffer the most on signing in reverberant conditions while STRAIGHT was the most robust.

An example of pitch tracking on speech using popular PDAs can be seen in Figure 4.2. All PDAs perform well except for CEP. This evaluation was not exhaustive and is included for descriptive purposes as opposed to evaluating pitch trackers on speech.

4.1.4 Pitch Tracking for Birds

Pitch extraction tools which have been proven to work for human speech and music may not work as well on birds. Bird vocalizations differ from speech in a number of ways. An important difference is the frequency range. Bird vocalizations tend to have a wider bandwidth and higher mean pitch than human vocalizations. In 2011 Tchernichovski released software called Sound Analysis Pro (SAP) [156] (also available as Matlab toolkit SAT). SAP calculates a number of features, one being $\boldsymbol{F_0}$ which is calculated using the YIN algorithm [42]. Mandelblat-Cerf and Fee [95] also used SAP for evaluating song imitation (also for zebra finches) where pitch again was a crucial feature. While zebra finch vocalizations may not be liable to pitch errors, YIN's

performance on other types of bird vocalizations is undocumented. Babacan et al. [5] discussed pitch tracking performance on singing sounds. While singing sounds are not identical to bird vocalizations they are more comparable than speech and bird song. Work in [5] showed YIN to have the lowest fine pitch error rate and 2nd lowest gross pitch error and F0 frame error after RAPT.

Early work leading to this work revealed that a number of pitch trackers do not match the spectrogram. An example of pitch tracking on bird song can be seen in Figure 4.3. The plot shows YIN and SWIPE' to accurately track the bird syllables while the rest of the PDAs do not. YIN performed slightly better voiced/unvoiced detection than SWIPE' here. Most parameters were the default settings except window size was set to 6.7 ms and frame rate to 1.7 ms. These values were chosen experimentally. Choice of window size accommodates the trade off between a large window which does not capture rapid pitch modulations, and a small window which does not capture as many periods. When investigating other PDAs, default parameters provided by authors were used. In some instances frequency range needed to be modified to allow pitch tracking of bird song which is a lot higher than humans. Further careful parameter selection for other PDAs may have brought their performance in line with YIN but Figure 4.3 shows YIN to be a good choice of pitch tracker for bird song with minimal initial tuning.

Based on these preliminary tests, the findings in [5], the use of YIN in SAP [157, 95], and its reputation in the speech community as a good pitch estimator for speech and music, YIN was chosen as the baseline pitch estimator for extracting pitch of bird recordings in this work. Although RAPT performed slightly better than YIN on singing sounds in [5], the pitch range of birds (>1666 Hz) is too large for RAPT [24], hence it was not chosen as the baseline in this thesis. YIN is based on the well-known autocorrelation method with a number of modifications [42]. Autocorrelation is very effective for pitch tracking, but some autocorrelation peaks suffer ambiguity, which leads to octave error or estimates too low in frequency [87]. Some bird vocalizations change frequency rapidly and over a wide range (1-5 kHz). Many syllables include extended frequency sweeps that sometimes exceed two octaves [96], which makes bird vocalizations prone to these types of errors.

Other pitch packages have been used within the bird community. Lachlan et al. developed Luscinia available at [84]. Luscinia is widely used in the community to analyze bird song [85]. It provides pitch extraction but requires supervision. The Luscinia GUI allows the user to select elements that require pitch tracking, but even when elements are carefully selected the exported pitch tracks may not always match the spectrogram. Meliza et al. developed Chirp [105], a tool which allows the user draw a mask on the spectrogram to improve pitch estimation. Time-frequency reassignment spectrographic analysis, harmonic template matching and Bayesian particle filtering are combined to produce pitch estimates [105]. While this method provides great pitch estimates when compared to the spectrogram, the masks need to be hand drawn on the spectrogram. Even with hand drawn masks the pitch can still be error prone.

The focus of this work is to present a pitch tracking algorithm suited to large scale processing

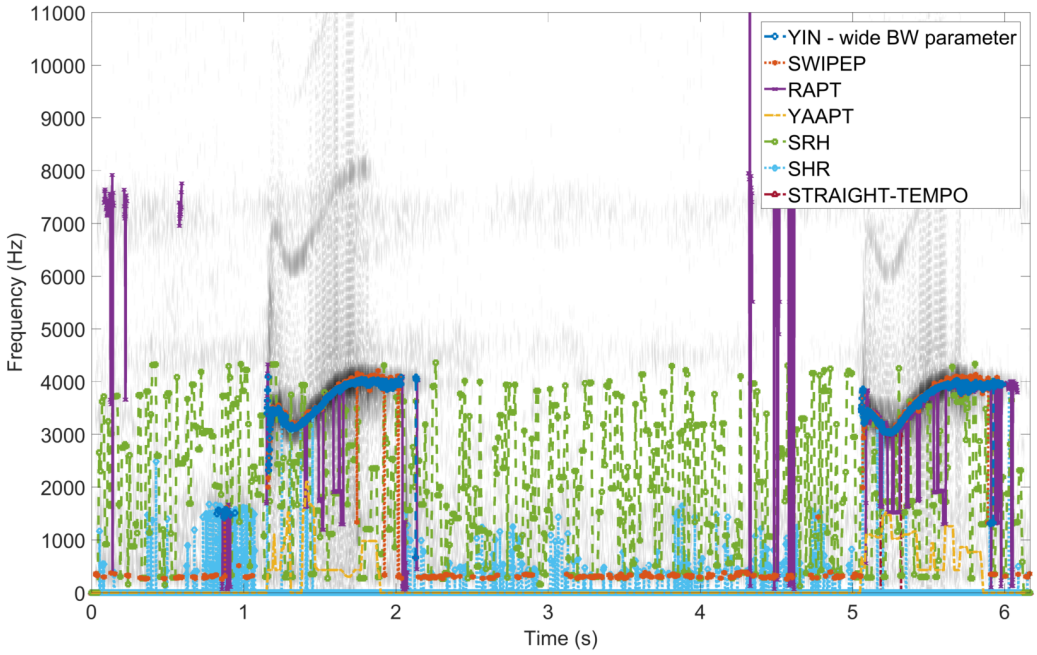


Figure 4.3: Pitch tracking on bird song of Eastern Wood Pewee (XC 1377 [166]).

of tracks that require no manual intervention other than initial tuning.

The improved system presented in this Chapter, referred to subsequently as YIN-bird, is described in Section 4.2. For the first time, a ground truth database of synthesized bird vocalizations with known pitch was developed to allow a quantitative evaluation of YIN-bird and is discussed in Section 4.3.2. The performance of YIN compared to YIN-bird is thus evaluated using the standard error metrics in the signal processing field (described in Section 4.3.4). Common types of bird vocalizations were presented in Section 2.3 in chapter 2 earlier. Using YIN-bird on a set of bird whistles improves gross pitch error from 1.67% to 0.58%, a relative improvement of 65.27%. For trills, the figure reduces from 6.29% to 2.31%, a relative improvement of 63.28%. Performance on other vocalization types are discussed in Section 4.4. Finally a qualitative analysis of more complex bird vocalizations and pitch tracking performance on these sounds are discussed in Section 4.5. Portions of this work was presented in [117] at INTERSPEECH 2016.

4.2 Tuning YIN to Bird Song

YIN processes audio data and outputs a pitch estimate. Parameters can be specified for each file, with a more accurate pitch estimate when parameters are carefully selected to match the input characteristics. One of the more sensitive input parameters is minimum frequency threshold ($F_{0_{min}}$). As bird vocalizations have a wider bandwidth than human speech, a single $F_{0_{min}}$ for all segments of the input file may not be suitable. The proposed system YIN-bird, determines

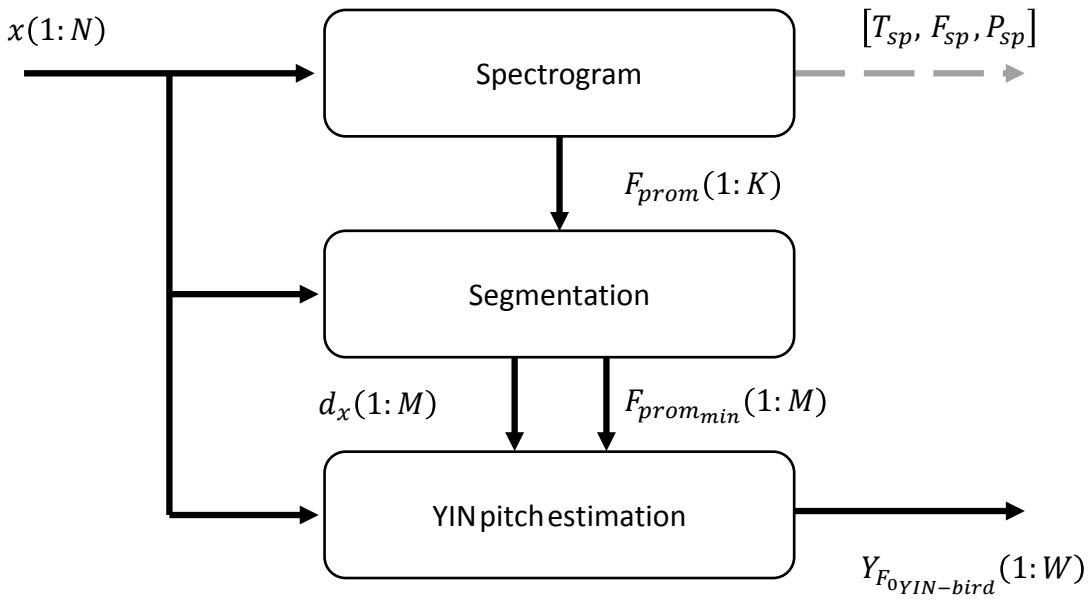
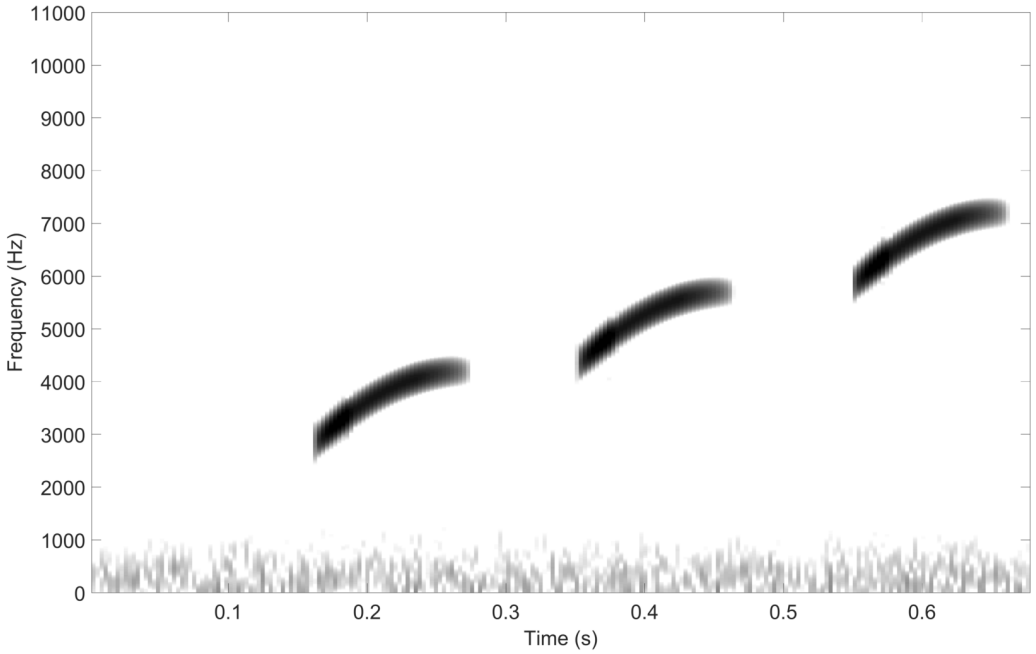


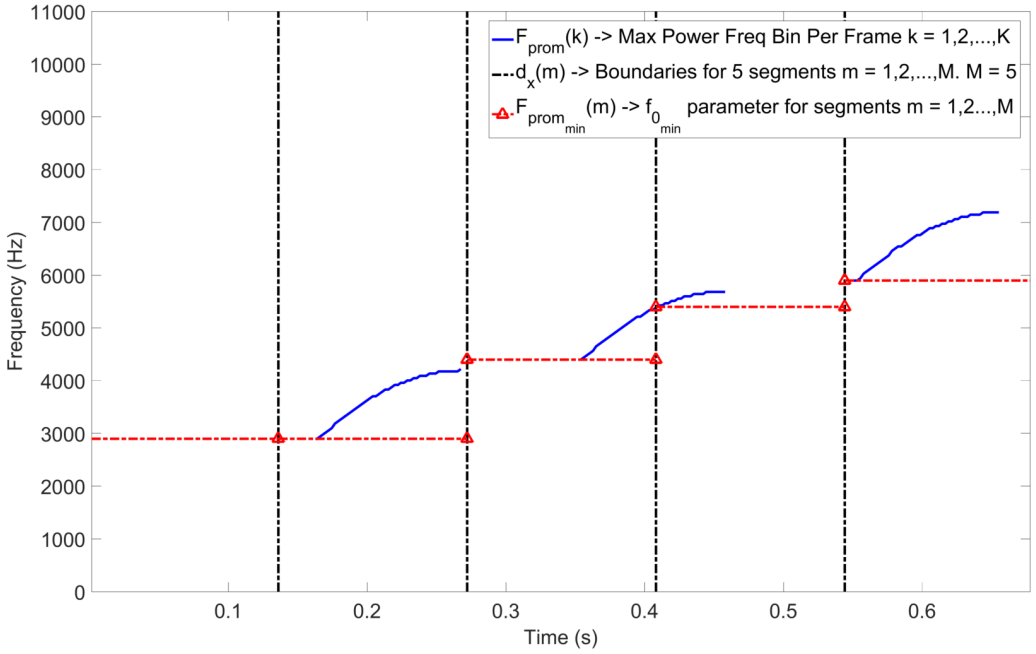
Figure 4.4: Block diagram of adaptive $F_{0_{min}}$ YIN (YIN-bird).

a suitable $F_{0_{min}}$ for each segment of a bird recording, through careful analysis of the input spectrogram. Using spectrogram information, each segment is assigned a $F_{0_{min}}$ parameter which leads to a more accurate pitch estimate for each input file. Maximum frequency parameter ($F_{0_{max}}$) had little influence on pitch estimates and was set to 0.4 x $\boldsymbol{F_s}$ where $\boldsymbol{F_s}$ is the sampling frequency of the input file. A block diagram of the system is shown in Figure 4.4.

Step 1: involves calculating the spectrogram parameters $[T_{sp}, F_{sp}, P_{sp}]$, of audio recording $x(1:N)$, where x is the input signal, N is the number of samples of the input signal, T_{sp} is the spectrogram frame time information, F_{sp} is the spectrogram frequency bins and P_{sp} is a matrix containing the power of each frequency bin at each time frame. Figure 4.5a shows a spectrogram of bird syllables. Using the power (dB) and frequency (Hz) information, a prominent frequency (i.e. frequency bin with most power) for each frame is selected automatically $F_{prom}(k)$ where $k = 1, ..., K$ and K is the number of frames in the spectrogram. Figure 4.5b has the prominent frequencies $F_{prom}(k)$ plotted with a continuous blue line. Information at frequencies of 200 Hz (value chosen by low pass filtering multiple bird recordings and checking for the presence of bird vocalizations) and below is assumed to be noise and is ignored. If the power of frame k's prominent frequency ($P_{F_{prom}}(k)$) is less than $mean(P_{F_{prom}}(1:K))$ for a given recording, frame k's prominent frequency is ignored (in practice, assigned 'Not a Number (NaN)' in Matlab) as it is most likely an unvoiced frame or a frame without vocalization (See Eq. 4.1).



(a) Spectrogram of bird whistles input to YIN-bird.



(b) Bird song prominent frequencies (continuous blue line), segment boundaries (vertical broken black line) and adaptive $F_{0_{min}}$ values (broken horizontal red line with circular markers) used by YIN-bird. F_s for this example was 22 kHz, therefore a segment of 3000 samples equates to 136 ms.

Figure 4.5: Elements of processing in YIN-bird.

$$If : P_{F_{prom}}(k) < mean(P_{F_{prom}}(1:K)) \quad then \;\; F_{prom}(k) = NaN \tag{4.1}$$

where k is the spectrogram frame number, $P_{F_{prom}}(k)$ is the power at the prominent frequency of frame k, $mean(P_{F_{prom}}(1:K))$ is the average of the power at each frame's most prominent frequency over a single recording and $F_{prom}(k)$ is the prominent frequency at frame k.

Step 2: segments the audio file into chunks specified by the user. The segment size is selected based on the bird corpus being used (small segment size gives slower execution). In this chapter each segment contains 3000 samples of input x (68 ms when $\boldsymbol{F_s}$ is 44.1 kHz). 3000 was chosen experimentally accounting for the trade off between slow processing for short segments and less frequent updates of $F_{0_{min}}$ for larger segments. All files in the data set were resampled to 44.1 kHz. Each segment is described as $d_x(m)$ where $m = 1, ..., M$ and M is the input number of samples (N) divided by 3000. Segments are shown divided by black broken vertical lines in Figure. 4.5b. Groups of prominent frequencies ($F_{prom}(1:K)$) are assigned to an appropriate $d_x(m)$. If K is 300 frames and M is 30 segments, then prominent frequencies values $F_{prom}(1:10)$ will be grouped in $d_x(1)$. The minimum F_{prom} in each $d_x(m)$ is $F_{prom_{min}}(m)$. $F_{prom_{min}}(m)$ for each frame is plotted with a broken horizontal line and circular red marker in Figure 4.5b. In Figure 4.5b the first two segments have the same value for $F_{0_{min}}$ ($F_{prom_{min}}(1) = F_{prom_{min}}(2)$). This is because if there is no vocalization within a segment, F_{prom_m} will take its nearest neighbor's value (or nearest neighbor with a value). If there are two nearest neighbors, previous values take precedence over posterior neighbor's value.

Step 3: involves processing the whole audio file (x) with YIN multiple times. This is purely to make timing information of YIN-bird's output consistent with YIN. The alternate option would have been to pass individual segments to YIN separately and concatenating $\boldsymbol{F_0}$ estimates, unfortunately this method was not straight forward and involved changing the way YIN runs. As processing time was not an issue, the whole audio file was passed to YIN each time. Each YIN estimation uses $F_{0_{min}}$ taken from $F_{prom_{min}}(m)$. $F_{prom_{min}}$ values are rounded to the nearest 100 Hz to reduce the number of times x is passed through YIN. If any two values in $F_{prom_{min}}$ are equal this reduces the number of times YIN is called from M to $M-1$. Once all the pitch estimates have been collected, each segment $d_x(m)$ is assigned its pitch estimate from YIN's output when $F_{0_{min}}$ equals $F_{prom_{min}}(m)$. Finally an output pitch vector from YIN-bird is concatenated, $Y_{F_{0_{YINbird}}}(1:W)$ (where W is number of pitch values, which is reliant on YIN's hop size (or frame rate, equal to 1.7 ms here) parameter).

To summarize YIN-bird contains a front-end to YIN that works out $F_{0_{min}}$ for each segment then calls YIN on that segment.

Table 4.1: Bird vocalization data.

Category	No. of examples	Length (min:sec)
Whistles & hoots	107	40 : 09
Trills	65	13 : 02
Nasals	63	12 : 32

4.3 Experimental Setup

These experiments have two main goals, to evaluate the accuracy of pitch tracking on different types of bird vocalizations, and to evaluate the benefits of using an adaptive $F_{0_{min}}$ parameter (YIN-bird).

4.3.1 Data

Examples of birds that produce sounds discussed in Section 2.3 are given at earbirding.com [121]. Recordings of these birds were downloaded from xeno-canto.org, a popular website dedicated to sharing bird sounds from around the world [166]. Recordings were pre-processed manually using Adobe Audition. This involved highlighting and deleting parts of the spectrogram that contained non-target vocalizations and silence. No other pre-processing or noise reduction was used. Thus the recordings varied in quality, as is typical of bird recordings taken in the wild. The data was grouped into 'Whistles & hoots', 'Trills' and 'Nasals', as introduced in Section 2.3 of chapter 2. The data is summarized in Table 4.1.

4.3.2 Synthesized Bird Sounds

Bird vocalizations contain voiced and unvoiced parts. One way to view pitch tracker accuracy is to superimpose the pitch tracker output on the spectrogram. A good estimate will show the pitch tracking the lowest spectral peak of the voiced parts. This is subjective. To overcome subjectivity, it is desirable to have a ground truth pitch to compare pitch tracker outputs to. Unfortunately no dataset with ground truth pitch for birds exists. The data in Table 4.1 inspired the creation of a synthesized bird sounds data set complete with ground truth pitch. Real recordings used to create a re-synthesized database. The synthesis system used here was taken from 'Spectral Modeling Synthesis (SMS) Tools' a python implementation for analysis, transformation and synthesis of musical sounds based on various spectral modeling approaches by Serra [142]. SMS contains a synthesis method called 'Sine plus Residual (SpR)', a method which uses peak detection.

The input to the system is real bird recordings. The aim is to synthesize the periodic parts of the vocalizations and save a ground truth frequency. The non-periodic residual ($Xr[k]$ in Figure 4.6)is then added to the synthesized periodic part to give a more natural sound. SpR

requires parameters, like number of sines (this was set to 8), which is the number of sine waves used in the synth phase, and the power threshold, which sets the amplitude threshold required for a given peak to be highlighted as periodic. Power threshold was usually set to -60 dB but this varied for some recordings as manual supervision was required to ensure optimal rejection of non-periodic peaks and accurate ground truth prediction. Every frame will have peaks, but only the high amplitude peaks (amplitude at frequency bin per frame) will be highlighted as peaks to be synthesized as sine waves. In theory high amplitude unvoiced peaks can be highlighted as periodic frames and low amplitude voiced frames can be missed as periodic frames for synthesis. In practice, only voiced parts will have a high concentrated amplitude at a frequency bin, whereas unvoiced sounds will have energy distributed across a wider bandwidth. Carefully choosing the threshold for different files prevents unvoiced parts being synthesized and being included in the ground truth pitch signal. Inevitably some gaps in the ground truth will appear due to amplitude fluctuations of bird recordings.

In summary, the output synth sound will contain synthesized sine waves and the original residual (non-periodic) audio added together in the frequency domain. The lowest sine track frequencies are saved as the ground truth. In some cases the ground truth needs manual correcting, e.g. when the fundamental is missing, the 2nd harmonic would be incorrectly picked as the fundamental frequency. This method is proposed as the optimal way to generate a ground truth without resorting to fully manual labeling of pitch.

The SpR code can be found at [143]. All files in the data set required resampling to 44.1 kHz to work with the SpR code. Each raw wave file of bird song ($x[n]$; $\boldsymbol{F_s}$ = 44.1 kHz) was passed through Serra's SpR system. A block diagram of the system can be seen in Figure 4.6. Input $x[n]$ was windowed using a Hamming window ($w[n]$). Taking the FFT of each window (Short Time Fourier Transform (STFT)) resulted in magnitude $|X[k]|$ and phase $\angle X[k]$. $|X[k]|$ and $\angle X[k]$ were passed to a peak detector. The amplitude, frequency and phase (A_p, f_p, P_p) of the peaks were then passed to the sinusoidal tracking block where sine tracks were identified. These periodic sine tracks were used by the sine spectral synthesis block to synthesize the periodic part of the bird vocalization. The non-periodic residual was found by subtracting the modeled periodic parts from the spectrum of the windowed input. The residual was added to the sine model to give a more realistic synthesized bird sound. The final synthesized sound ($y[n]$) was constructed by adding the residual spectrum ($X_r[k]$) to the periodic sine spectral synthesis output ($Yh[k]$) and calculating it's Inverse Fast Fourier Transform (IFFT) to give $y_w[n]$. This windowed signal passed through an overlap and add block to get the synthesized output signal ($y[n]$). The ground truth pitch ($g[n]$) was identified as the lowest frequency peak of the sine model over time. For unvoiced regions, $g[n]$ was assigned a value of 'NaN'. The parameters used for window and hop size were the same as YIN-bird so the ground truth signal would be the same length as the YIN-bird output.

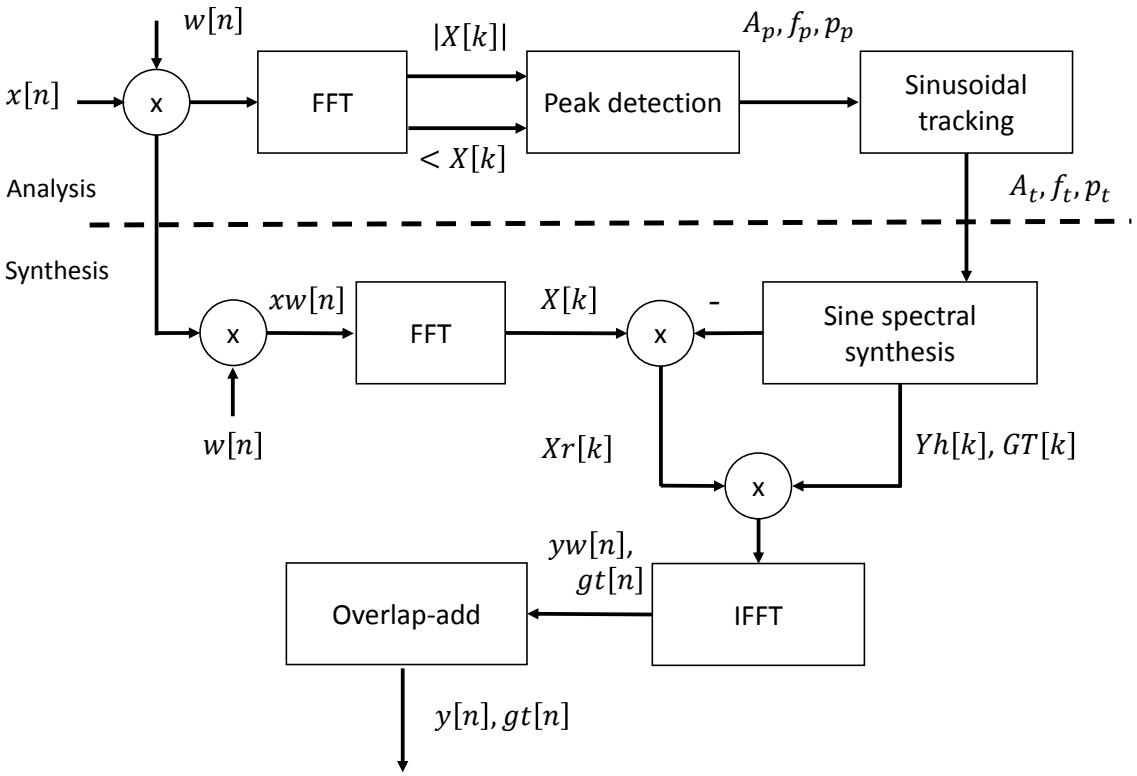


Figure 4.6: Block diagram of SpR system used to synthesis the bird data with ground truth pitch [142].

4.3.3 Listener Tests

Listener tests were performed online to evaluate how well the synthetic sounds match the original recordings. Listeners were asked to listen to original recordings of bird vocalizations followed by the synthesized version and compare the pair of audio clips on a scale of {-3, ..., 3} (see Table 4.2 for scale description). The scale was influenced by work in [136], where listeners were asked to evaluate the speaker recognizability of synthetic speech using a similar scale. Three training examples were included with a recommended score revealed after listeners gave their answer. Participants were asked "how are you listening?". 36% used over-ear headphones, 22% in-ear headphones, 21% laptop speakers, 7% HQ external speakers and 14% regular external speakers. Of 23 respondents, 13 described themselves as 'Expert' listeners, 15 as 'Intermediate' listeners and 1 as 'beginner' listener with regards to their understanding of bird song. There were 26 synthetic examples tested. A screenshot of the survey can be seen in Figure 4.7. The survey was designed on surveygizmo.com and remains available at [115]. On a 6 point scale, worst to best from {-3, ..., 3}, the average score of the 23 listeners was 2.17 which describes the synthetic sound as "Sounds very much like original, could be fooled into thinking it is a real bird". This

Table 4.2: Listener testing evaluation scale.

Evaluation	Rate	Description
Very different	-3	Doesn't sound like original, clear its a synth version
Fairly different	-2	Sounds slightly different than original, most likely a synth
Little different	-1	Sounds like original, most likely a synth
Fair	0	Sounds like original, might be a synth
Little similar	1	Sounds like original, unsure if a synth
Fairly similar	2	Sounds very much like original, could be fooled into thinking it is a real bird
Very similar	3	Sounds identical to original, confident its a real bird recording

test was to clarify that the synthetic sounds are similar enough to the original recordings, that they can be used in the pitch estimation experiments here. A more exhaustive listener test would have to be performed if this chapter was presenting a novel synthetic system.

4.3.4 Error Metrics

Performance of the two pitch tracking systems was assessed using four standard error metrics [171, 5].

- **Gross Pitch Error (GPE)** is the percentage of frames for which the absolute pitch error is higher than a certain threshold. For speech this threshold is usually 20%. As bird vocalizations tend to have higher pitch the threshold was reduced to 10%. Only frames considered voiced by both the pitch tracker and ground truth were included in this calculation.
- **Fine Pitch Error (FPE)** is the standard deviation of the absolute error in Hz. Frames that have gross pitch errors were excluded. Only frames with ground truth and YIN estimates being voiced were used to calculate FPE.
- **Voicing Decision Error (VDE)** is the percentage of frames for which an incorrect voiced/unvoiced decision is made.
- **F0 Frame Error (FFE)** is the percentage of frames where either a GPE or VDE is observed.

4.3.5 Experiment Parameters

The commonly used YIN system was compared with YIN-bird. For YIN, parameters wide enough to accommodate all bird vocalizations were used. $F_{0_{min}}$ was 500 Hz, window size was 6.7 ms, hop size was 1.7 ms (approximately 75% overlap) and quality was 'good' which means

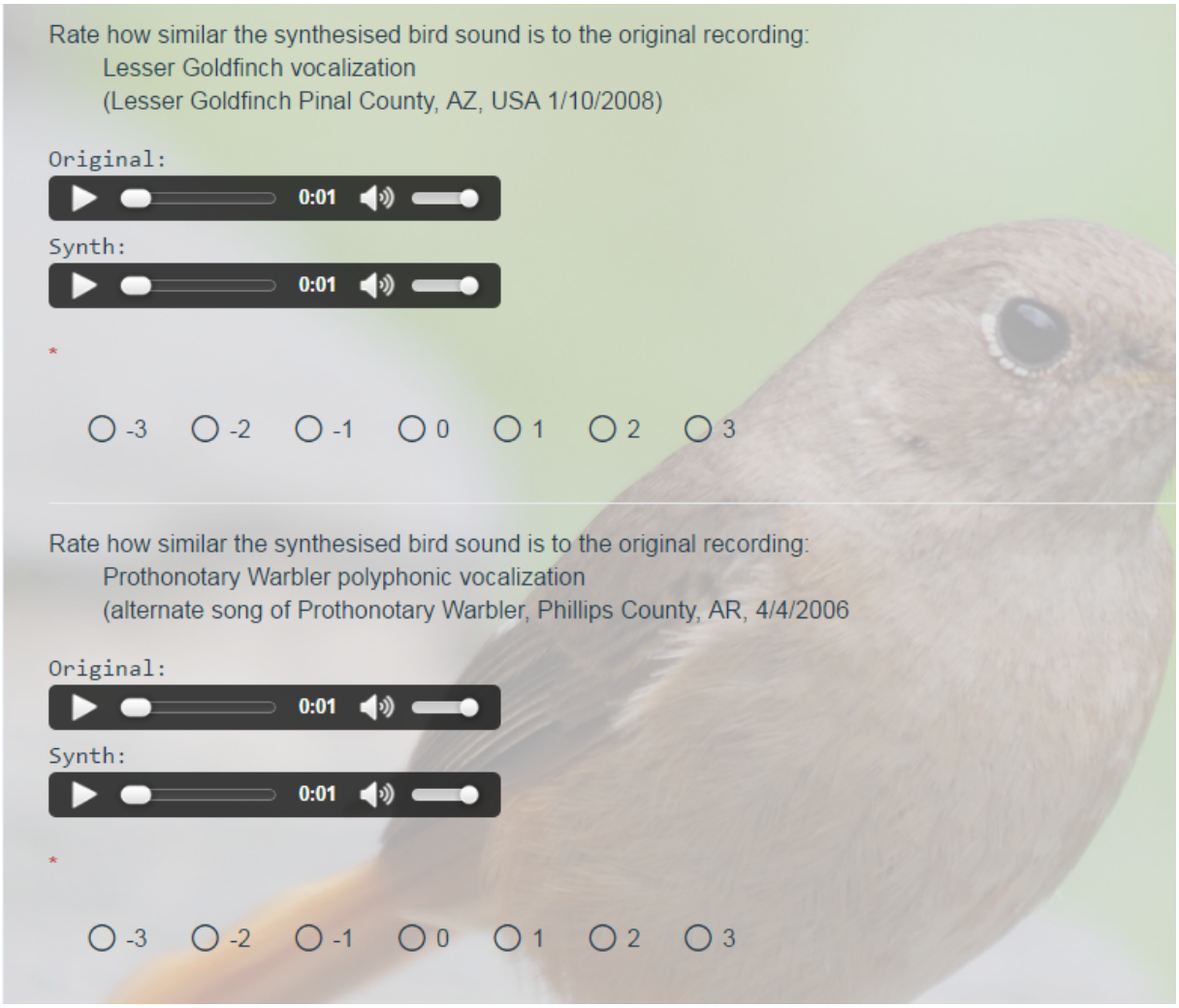


Figure 4.7: A screenshot of listener test used to evaluate similarity between synthetic bird song and original recordings [115].

estimates with aperiodic value of less than 2 x 0.2 (1 x 0.2 for 'best', convention used in YIN code) were considered voiced. For the trills the window size was reduced to 2 ms for increased time resolution, as pitch changes more rapidly for these type of sounds.

YIN-bird used the same window sizes as used with YIN above. No $F_{0_{min}}$ needed to be specified. The buffer size was set to 3000 samples, meaning that for every 3000 samples (68 ms) of the input audio file there would be a new value for $F_{0_{min}}$ parameter.

4.4 Quantitative Results

Pitch estimates using YIN, with parameters mentioned in Section 4.3.5, were compared to ground truth pitch ($g[n]$) in Hz from the synthesized values. Pitch estimates using YIN-bird were also

Table 4.3: Error rates using YIN & YIN-bird.

	GPE (%)	FPE (Hz)	VDE (%)	FFE (%)
Whistles:				
YIN	1.67	40.97	25.72	26.37
YIN-bird	**0.58**	**39.41**	**23.68**	**24.09**
Trills:				
YIN	6.29	88.89	37.93	41.12
YIN-bird	**2.31**	**63.75**	**35.61**	**36.78**
Nasals:				
YIN	31.00	42.60	33.28	48.94
YIN-bird	**6.21**	**58.67**	**32.69**	**35.60**

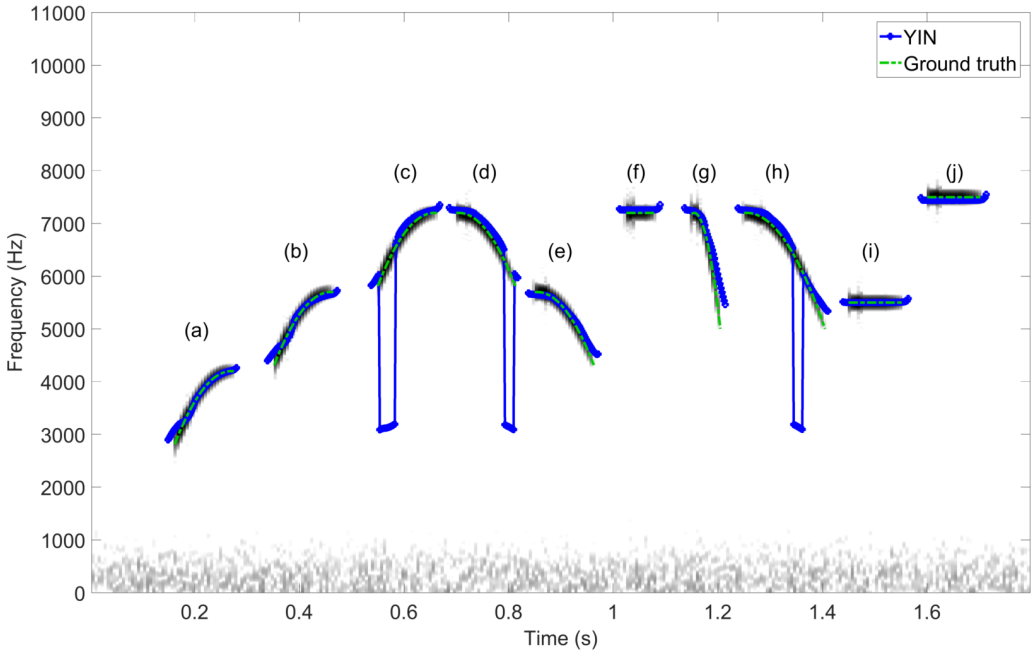
Table 4.4: Absolute and relative improvement in error between YIN and YIN-bird.

	GPE (%)	FPE (Hz)	VDE (%)	FFE (%)	P-value of GPE
Whistles:					
Abs	1.09	1.56	2.04	2.28	0.0831
Rel	**65.27**	**3.81**	**7.93**	**8.64**	
Trills:					
Abs	3.98	25.14	2.32	4.34	9.9811e-10
Rel	**63.28**	**28.28**	**6.11**	**10.55**	
Nasals:					
Abs	24.79	16.07	0.59	13.34	1.844e-06
Rel	**79.97**	**37.72**	**1.77**	**27.26**	

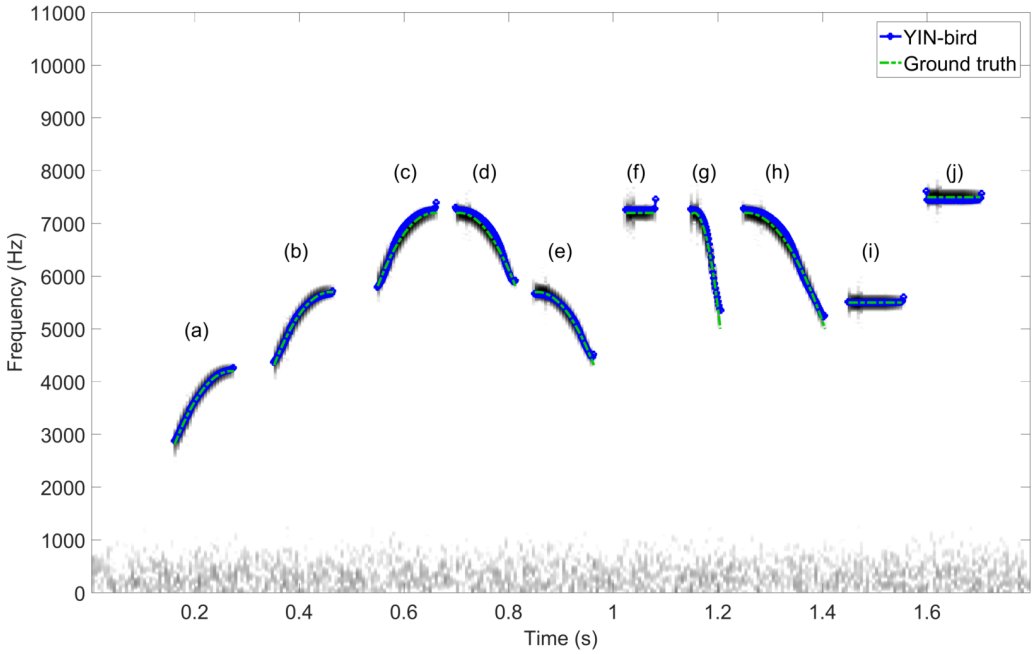
compared to the same ground truth. The results are shown in Table 4.3.

Table 4.4 presents the absolute and relative improvement in GPE, FPE, VDE and FFE using YIN-bird compared to YIN. When using YIN-bird for whistles, the GPE score shows an absolute improvement of 1.09% and a relative improvement of 65.27%. For trills, the absolute improvement in GPE is 3.98% and relative improvement is 63.28%. The GPE improvement for whistles is significant at the 10% significance level while trills and nasals are significant at the 1% significance level.

Typical YIN and YIN-bird performance on whistle sounds is shown in Figure 4.8. This shows how YIN performs on synthetic bird syllables created in Matlab. Note syllables (c), (d) & (h) experience octave errors or errors too low in Figure 4.8a ($g[n]$ is plotted with a broken green line and the YIN pitch estimate is plotted with circular blue markers. The same errors are observed using SAP [156]. Similar errors are produced by real data. These errors are corrected in Figure 4.8b, where pitch values obtained YIN-bird are plotted.



(a) YIN - whistles.



(b) YIN-bird - whistles.

Figure 4.8: Performance Comparison between YIN and YIN-bird on synthetic bird whistles. Pitch is plotted with circular markers in blue and ground truth with broken line in green.

Fine pitch error and voice detection error are included in Table 4.3 to show that the addition of an adaptive $F_{0_{min}}$ in YIN-bird does not diminish FPE and VDE. YIN-bird reduces 'pitch being too low' errors exclusively so VDE will not improve directly with YIN-bird. As FFE combines GPE and VDE, it can be used as an overall measure of pitch estimation performance [171, 5]. For whistles and trills the FFE absolute improvement is 2.28% & 4.34% respectively and relative improvement is 8.64% & 10.55%. An example of pitch tracking improvement for trills can be seen in Figure 4.9a & 4.9b.

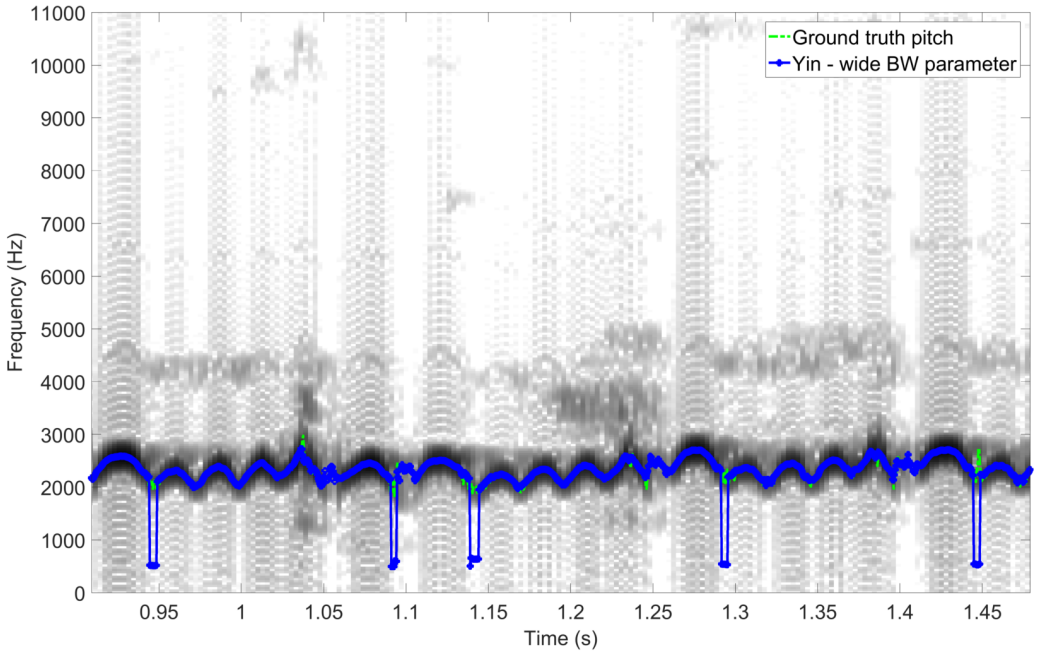
YIN-bird has reduced GPE and FPE for the ground truth data set of whistles and trills. Not all bird sounds are a single tone. Nasal sounds contain many harmonics. Pitch tracking on nasal sounds with multiple partials is a challenge, especially when $\boldsymbol{F_0}$ is missing, as is possible. Although the GPE results can be presented as an improvement for nasals, Figure 4.10a & 4.10b show how the pitch estimations jump between bands for both YIN and YIN-bird for nasal sounds. YIN-bird tends to identify the pitch as the strongest partial instead of $\boldsymbol{F_0}$. If $\boldsymbol{F_0}$ is weak or missing, YIN-bird will set $F_{0_{min}}$ to the prominent partial thus estimating $\boldsymbol{F_0}$ to be the prominent harmonic rather than the absent $\boldsymbol{F_0}$. YIN sometimes identifies a weak $\boldsymbol{F_0}$ but other times estimates a higher partial. The correct ground truth for nasals is also difficult to establish. The synthesis system here is prone to suggesting the strongest partial to be the ground truth as opposed to the true $\boldsymbol{F_0}$ and manual ground truth corrections were required for some nasal examples. This is why nasal results are presented with caution.

It is worth noting that the validity of the comparison to ground truth heavily relies on the accuracy of the ground truth used. Section 4.3.2 discussed the establishment of the ground truth pitch for the synthesized songs and admitted that on some occasions errors may be present (e.g. the highlighted areas of Figure 4.11). The only alternative would be hand labeling combined with expert listening. In the absence of such a dataset, it is felt the synthesized dataset represents the best possible trade-off.

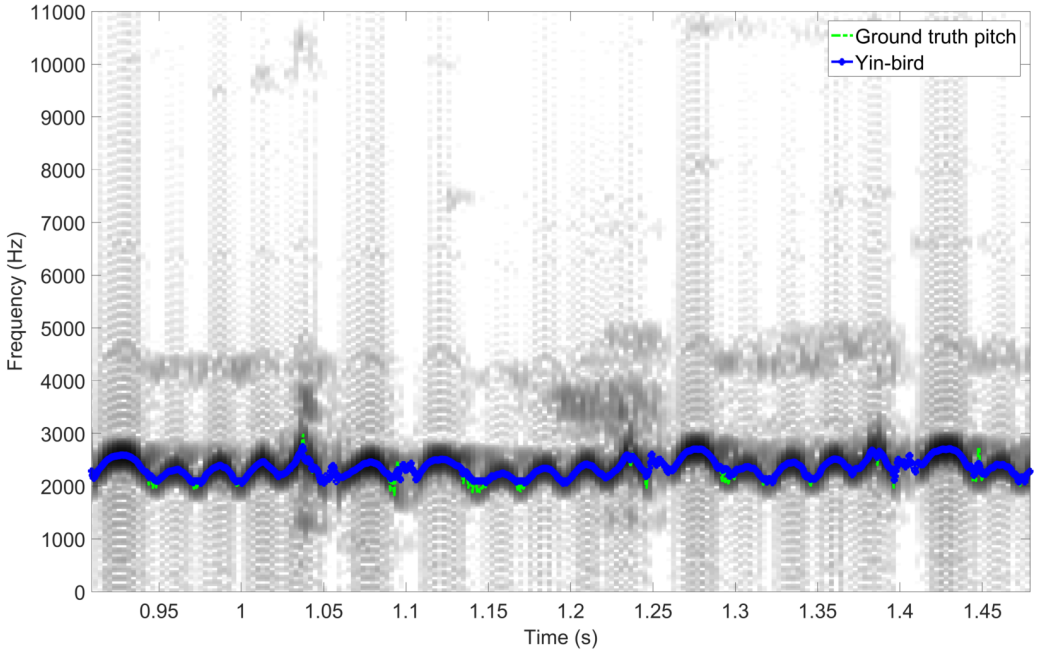
Where energy in partials is higher than that at the fundamental, timbre of the sound will change, how this affects the birds perception is not known. Perhaps to some birds, quality is more important than pitch? If $\boldsymbol{F_0}$ is missing, the interval between harmonics could be used to calculate $\boldsymbol{F_0}$, but if harmonics are mistuned or missing then this method will fail also [76].

The value of YIN-bird lies not only in this performance improvement, but also in the fully automatic processing of bird song. Nonetheless accurate pitch estimation and harmonic identification of harmonic/nasal sounds would benefit bird song research and needs to be made more consistent in further work.

A method which may overcome the problem of non-periodic peaks is to analyze the spectral shape to identify sinusoidal components in the short-time spectrum. This work was presented by Janĉoviĉ et al. in [64] & [67], and is listed in future work of chapter 7.

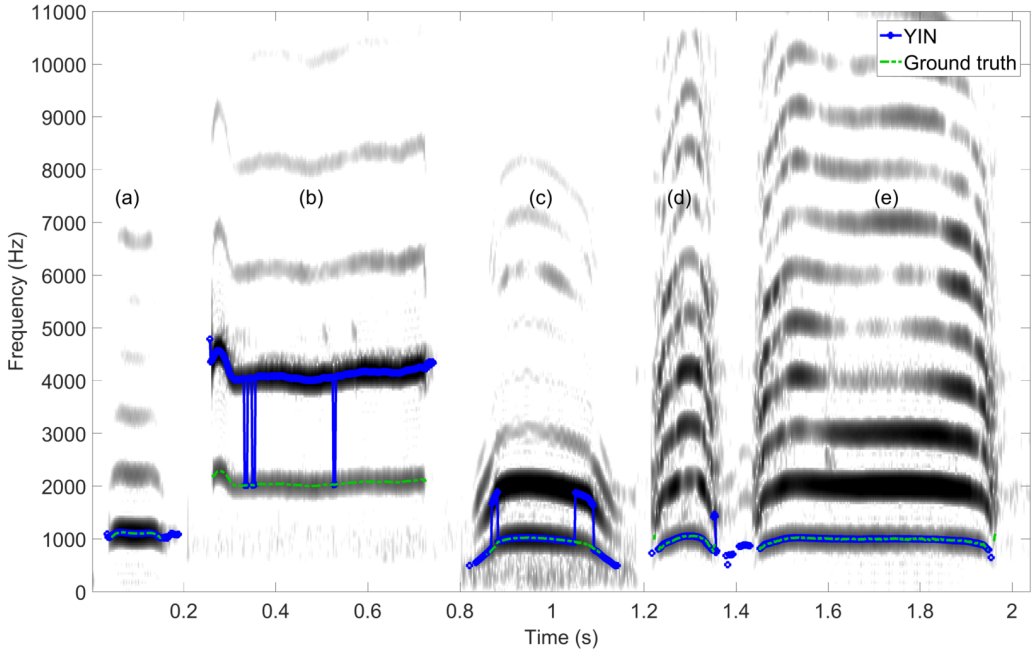


(a) YIN - trills.

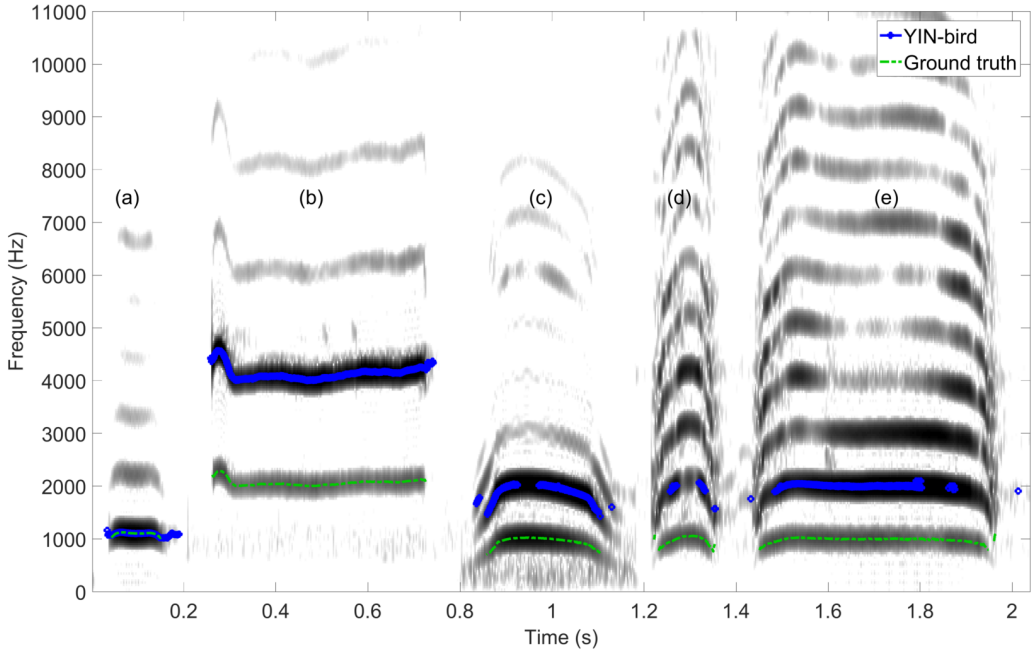


(b) YIN-bird - trills.

Figure 4.9: Performance Comparison between YIN and YIN-bird on a real bird trill example from a Ash throated Flycatcher XC286838 [166]. Pitch is plotted with circular markers in blue and ground truth with broken green line.



(a) YIN - nasal sounds.



(b) YIN-bird - nasal sounds.

Figure 4.10: YIN and YIN-bird pitch estimation on nasal sounds. Pitch is plotted with circular markers in blue and ground truth with a broken line in green.

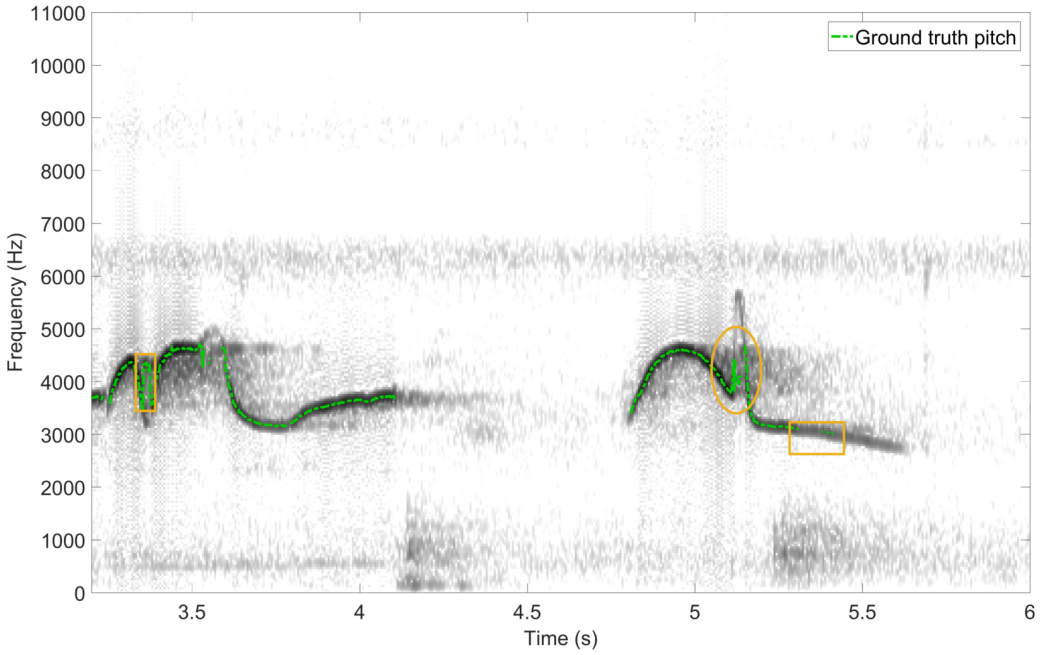


Figure 4.11: Example of ground truth pitch errors due to reverb and amplitude variation of Eastern Wood Pewee synth example (XC 7704 [166]). Note the ground truth errors inside the rectangles and circle.

4.5 Qualitative Analysis

Syllable types that could be synthesized with accurate ground truth pitch were used to quantitatively evaluate YIN-bird performance. Pitch performance on syllable types that posed problems when calculating ground truth pitch are discussed qualitatively here. These sounds can still be successfully synthesized even though determining a ground truth pitch was not possible.

4.5.1 Harmonic (or Nasal) pitch tracking

As mentioned in Section 4.4, not all bird sounds are a single tone. Harmonic (or Nasal) sounds with strong energy at $\boldsymbol{F_0}$ are straightforward for PDAs, just like harmonics in speech do not not diminish PDAs' performance for humans. When the 2$^{\text{nd}}$ harmonic (2 x $\boldsymbol{F_0}$, sometimes referred to as the 1$^{\text{st}}$ overtone or 2$^{\text{nd}}$ partial) is stronger than $\boldsymbol{F_0}$, the PDA jumps between $\boldsymbol{F_0}$ and the prominent partial as shown in Figure 4.12a at syllable 2, 3 & 5 using YIN. YIN-bird will set $f_{0_{min}}$ estimate to the 2$^{\text{nd}}$ harmonic and thus will track a single partial but incorrectly return the 2$^{\text{nd}}$ harmonic as $\boldsymbol{F_0}$ as shown in Figure 4.12b at syllable 2, 3 & 5. An example of missing $\boldsymbol{F_0}$ can be seen in Figure 4.12 at syllable 1. In Figure 4.12, $\boldsymbol{F_0}$ is missing from the first syllable and appears faintly for the other syllables. The 2$^{\text{nd}}$ harmonic is the prominent partial throughout the phrase in Figure 4.12. YIN pitch estimates jumps between $\boldsymbol{F_0}$ and 2$^{\text{nd}}$ harmonic in Figure 4.12a, while

YIN-bird tracks the 2nd harmonic in Figure 4.12b.

The problem of missing fundamental, missing harmonic and stronger higher order harmonics needs to be addressed as neither YIN or YIN-bird offer solutions to these problems. The advantage of YIN-bird is it will at least identify the prominent frequency partial and is less likely to jump between bands. Although YIN-bird is less likely to jump between partials while pitch tracking, when the prominent frequency changes throughout a vocalization the pitch estimation will jump as seen in Figure 4.13.

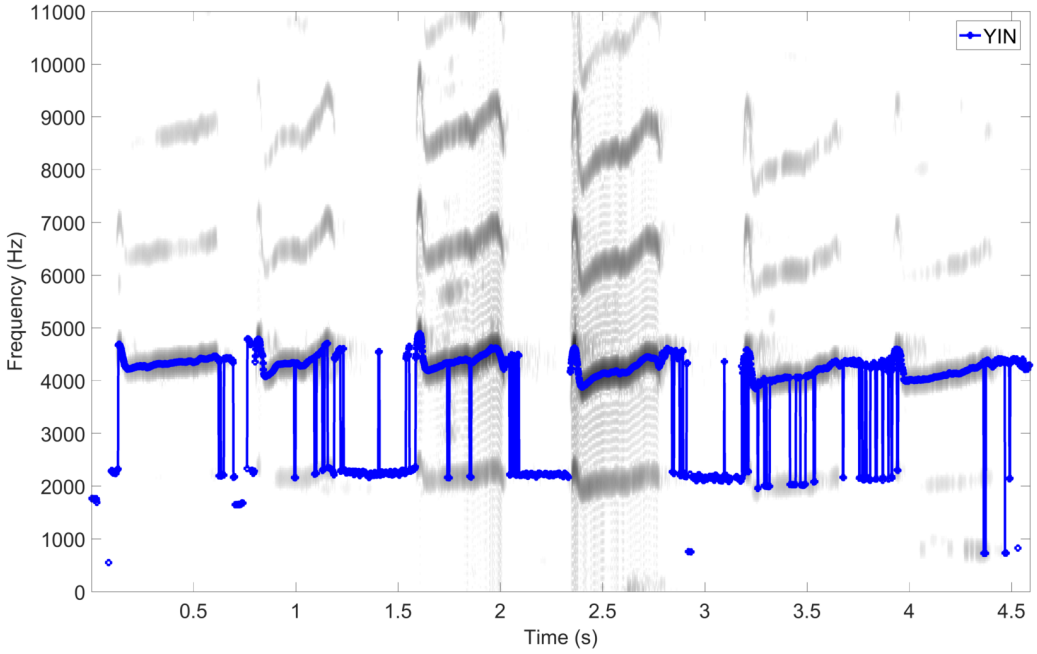
4.5.2 Two-voiced

Two-voiced sounds were introduced and two-voiced production explained in Section 2.3 of chapter 2. YIN-bird pitch tracking performance on a two-voiced example can be seen in Figure 4.14. PDAs get confused between the two tracks and output estimates that try and find an estimate average or common denominator between the two pitch tracks. This ultimately does not give you the pitch of the signal.

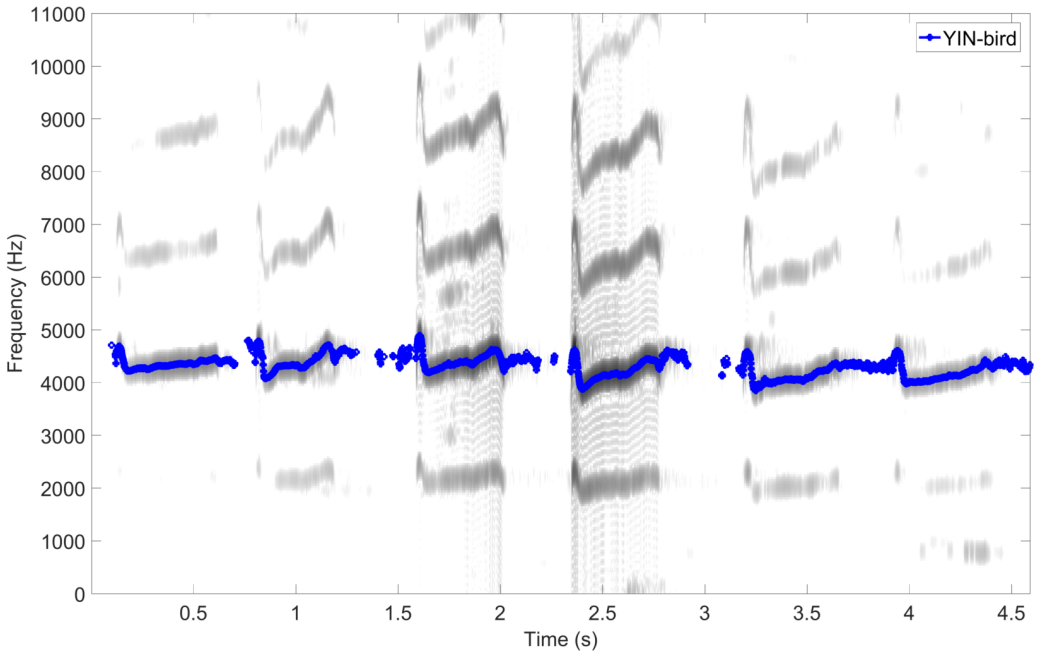
Sometimes two-voiced sounds can resemble harmonic sounds. Nowicki and Capranica [111] showed that a Chickadee's sound which resembled a harmonic sound was actually produced by heterodyne frequencies resulting in cross-modulation between the two syringeal sides. Pitch tracking of this phenomenon can be seen in Figure 4.15. Note the pitch estimate is at $\boldsymbol{F_0}$ sometimes even though it is missing from the vocalization. Pitch estimation using YIN-bird on this example tracks the partial at 3 kHz as the pitch (i.e. the prominent partial). In a true harmonic series $\boldsymbol{F_0}$ is usually the lowest and loudest partial but here a Red-breasted Nuthatch uses both sides of the syrinx simultaneously but not independently, instead vibrating sources are coupled together to create what looks like a harmonic series but technically isn't because $\boldsymbol{F_0}$ is missing and the prominent frequencies are around 3 kHz [121].

4.5.3 Noise

Not all songs and calls are tonal and smooth. Hissing, noisy and croaky sounds lead to messy spectrograms with sound diffused across the spectrum [96]. Sometimes a sound will start off voiced and ends up noisy. Noisy sounds have no pitch as they are non-periodic signals. Some bird sounds occur when noise passes through their narrow band filter vocal tract which appears as narrow band noise on a spectrogram. YIN and YIN-bird sometimes outputs a pitch when its an unvoiced vocalization as seen in Figure 4.16 where a Red-tailed Hawk makes a sound that starts off voiced and becomes noisy from 0.5 seconds to end (i.e. pitch estimate should stop once vocalization becomes unvoiced). An improvement to voiced/unvoiced detection for bird song is another sought after tool.



(a) YIN pitch.



(b) YIN-bird pitch.

Figure 4.12: Pitch tracking on a real recording of Killdeer song with missing F_0 (XC 173884 [166]).

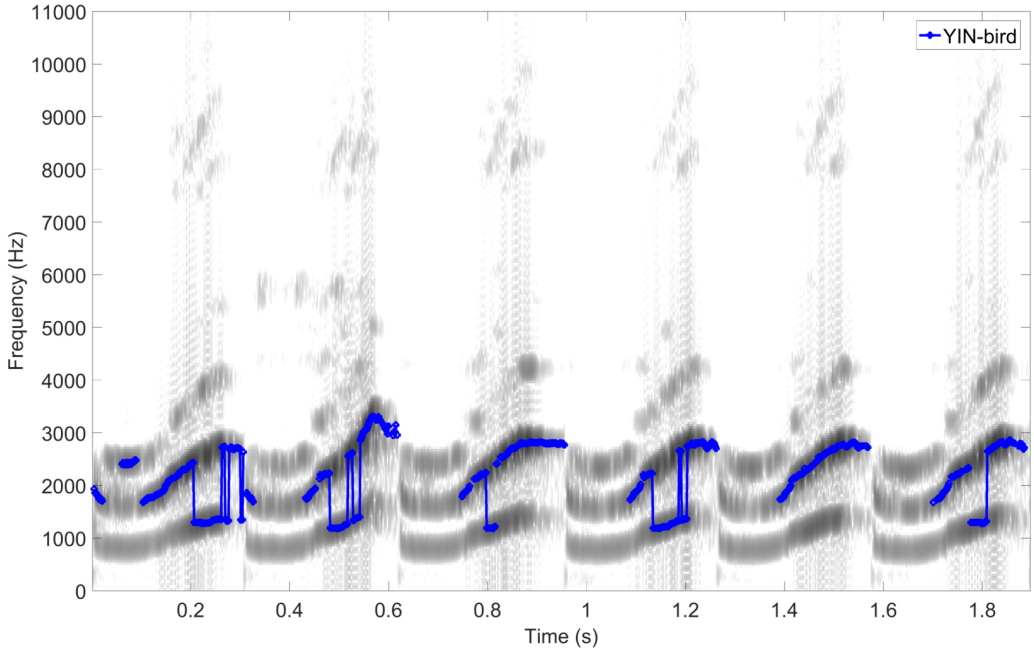


Figure 4.13: YIN-bird performance on a real recording of Sora where the prominent frequency changes between F_0 and 2nd harmonic (XC 292447 [166]).

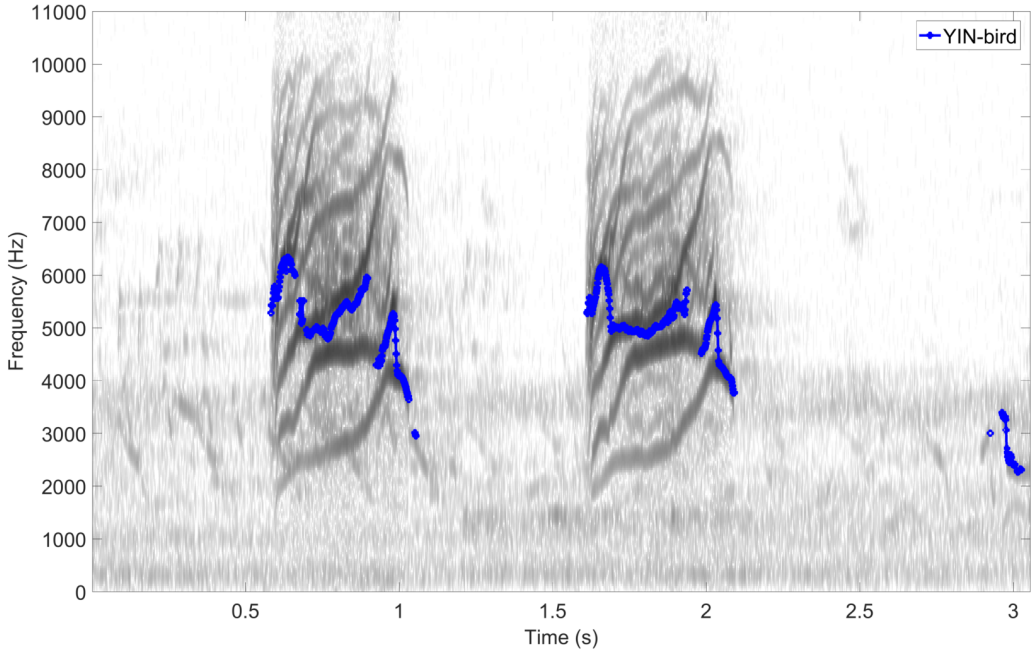


Figure 4.14: Typical performance of YIN-bird with two-voiced song shown on a real recording of Pine Siskin song (XC 97358 [166]).

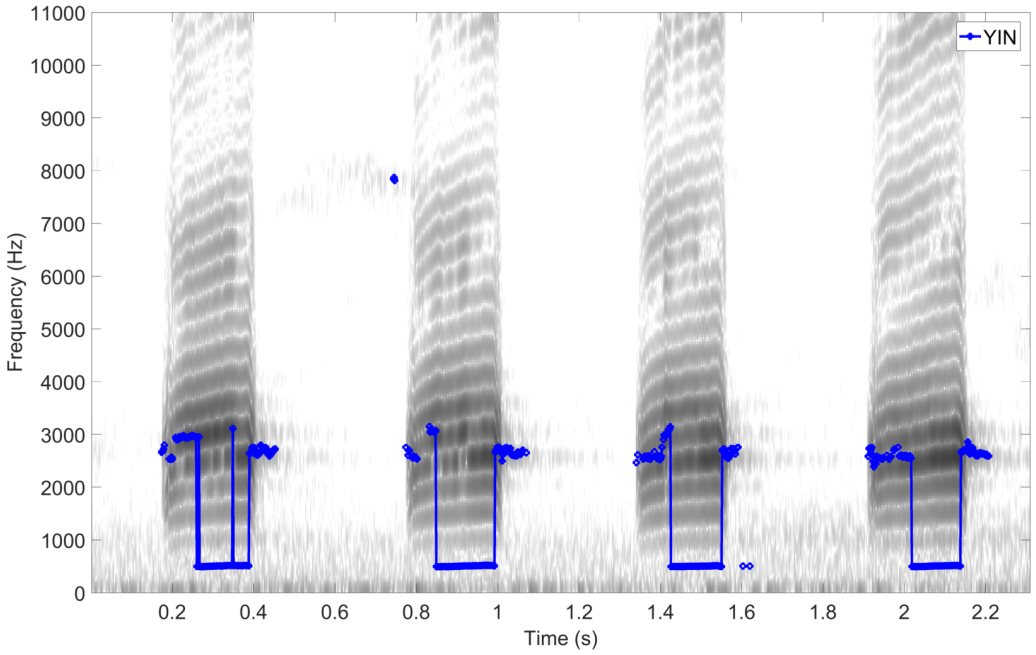


Figure 4.15: Red-breasted Nuthatch song. Harmonic like sound produced by syringeal coupling [121, 111] (XC 76418 [166])

.

4.5.4 Biphonation

Zollinger et al. in [177] discussed biphonation or two sounds coming from the one source (one side of the syrinx). Biphonatation is different to two voiced sounds. Biphonatation refers to a signal produced from one side of the syrinx being modulated to produce more than one simultaneous tones. Biphonation appears as a tone with side-bands. Amplitude modulation can create side-bands, one on each side of the carrier frequency [96]. Some Mockingbirds' vocalizations exhibit a phenomenon known as dual biphonation, where each side of the syrinx produces a tone and each tone is modulated by an unrelated lower frequency (possibly amplitude modulating frequency) resulting in a pair of side bands for each tone giving six lines on the spectrogram three for each side of the syrinx (combination of two voiced and biphonation phenomena). Biphonation was discussed in great detail in [177] but tracking pitch of such an occurrence is not trialed yet.

If a tone is frequency-modulated it can be described in two ways, as a warble or as a complex tone. Both are valid and what you see on a spectrogram depends on the analyzing bandwidth. A warble appears on a spectrogram with high time resolution as a tone fluctuating in frequency. A complex tone, on a spectrogram with high-frequency resolution, you see side-bands spaced at the modulation frequency, above and below the tone. The more side-bands there are, the more emphatic the warble sounds [96].

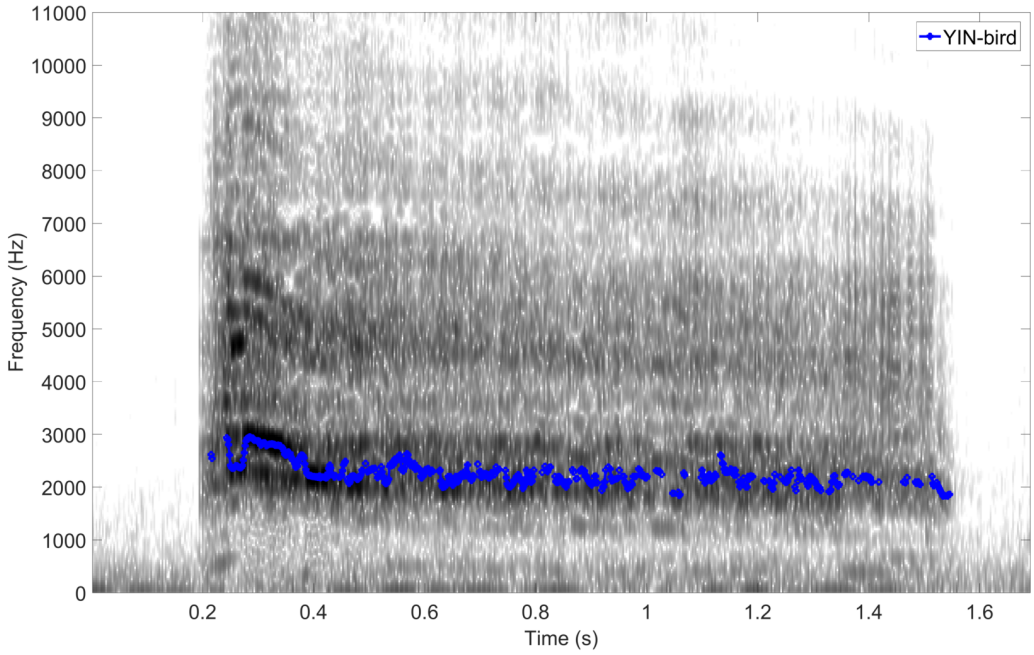


Figure 4.16: Red-tailed Hawk song that starts off voiced and becomes noisy. YIN-bird continues to output a pitch even after the syllable becomes a croaky unvoiced sound from 0.5 seconds onwards on this real recording example [121].

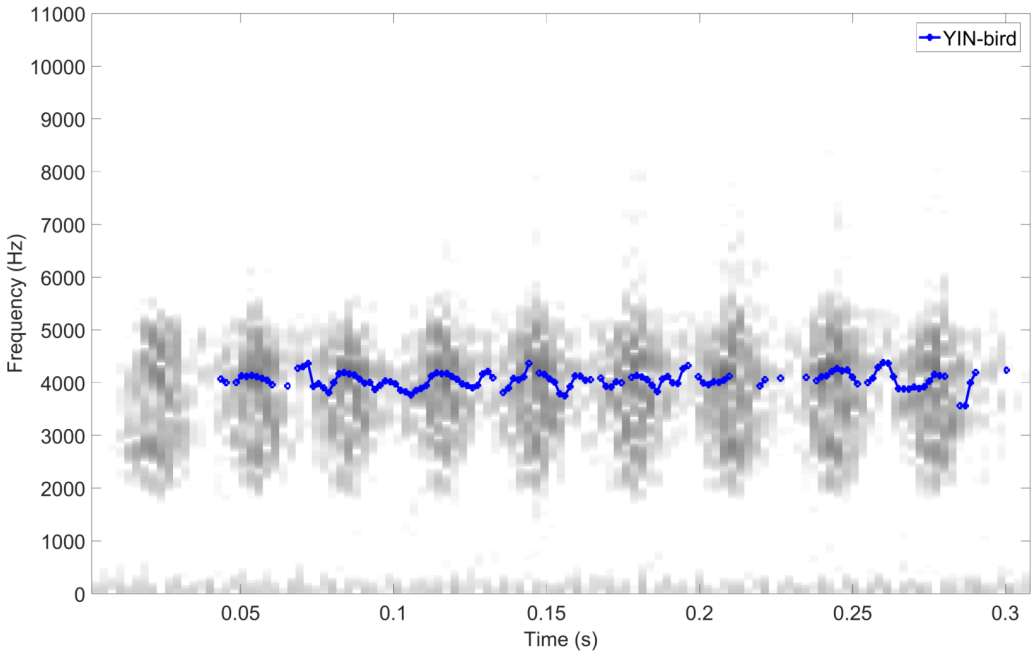
4.5.5 Birds with complex vocalizations

Nightingales, Sedge Warblers, Sky larks and Wood Warblers are just some of the birds that produce complex vocalizations. Figure 4.17 shows four complex phrases produced by Nightingales. Figure 4.17a shows a series of rapid harsh sounds which sounds like amplitude modulated narrow band noise. Figure 4.17b shows three phrases repeated. The first syllable is a fast whistle with a 2^{nd} harmonic. The second syllable is made up of multiple partials but not a harmonic series. This is more than likely syringeal coupling or two-voiced biphonation and sounds harsh.

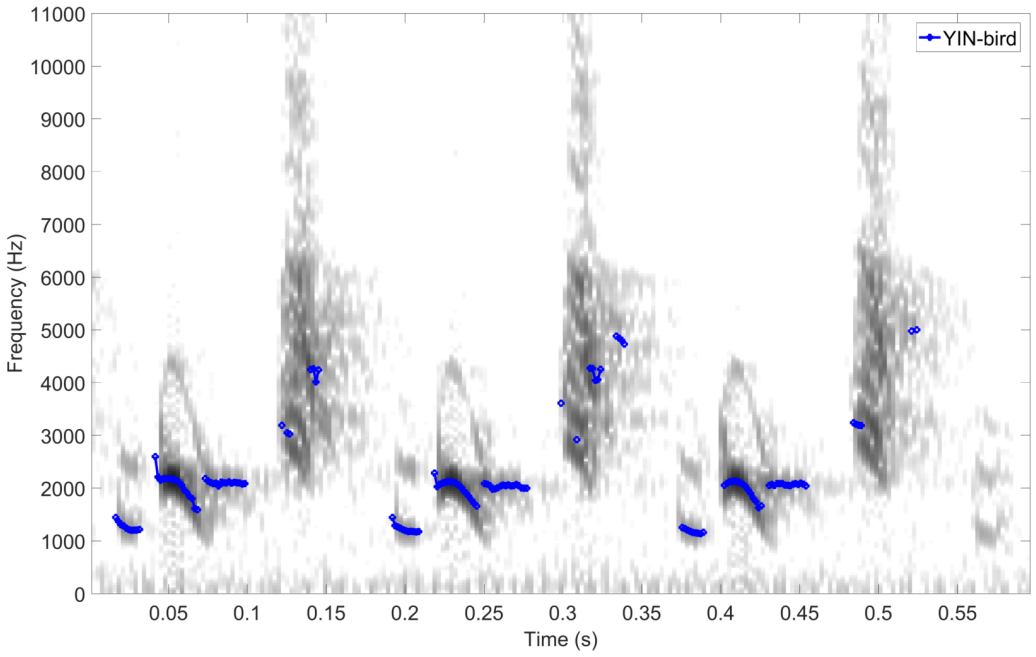
Examples of Sedge Warbler and Sky Lark vocalizations can be seen in Figure 4.18. The syllables in Figure 4.18a shows what appears to be syringeal coupling. Greater amplitude at the higher partials make them harsh sounding. Figure 4.18b shows rapid warble sounds. The time resolution had to be refined in order to display these syllables.

These examples show the complexity of nightingale, sedge warbler and sky lark song. Each syllable in isolation may not be complicated but the speed at which they vocalize and how rapidly the syllable type changes from whistles to trills or two-voiced leads to their overall complexity.

Throughout this Section, an in depth description of the challenges in tracking pitch of bird song is presented. Much work remains to be done before pitch tracking of complex song, like the song of Nightingale, Sedge Warbler and Sky Lark, is performed automatically and accurately.

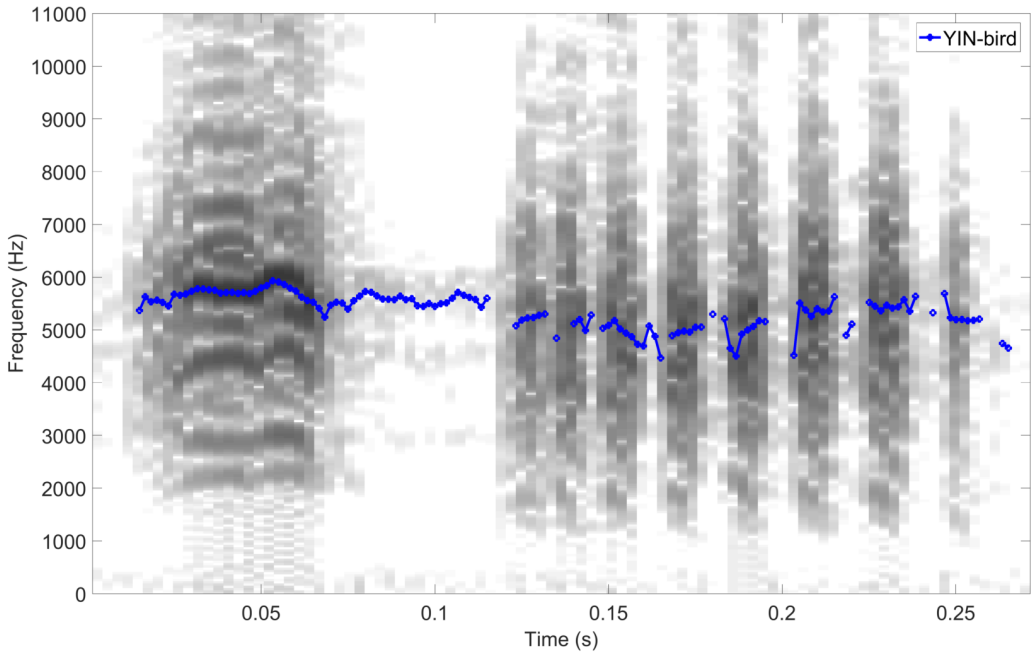


(a) Nightingale vocalization.

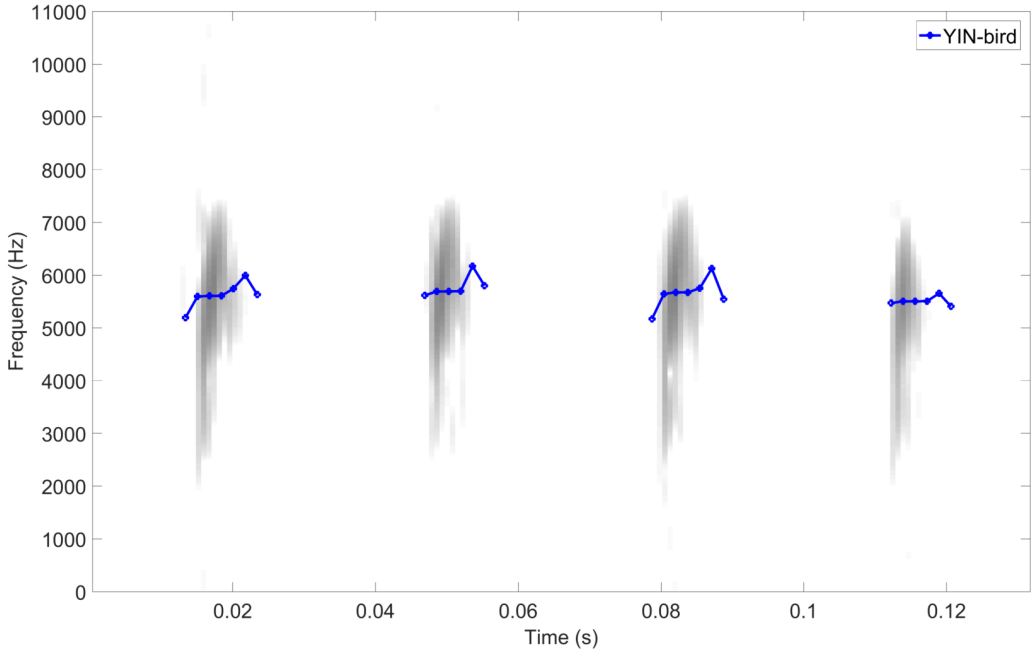


(b) Nightingale vocalization.

Figure 4.17: Real recordings of complex vocalizations of Nightingale [166].



(a) Sedge Warbler vocalization.



(b) Sky Lark vocalization.

Figure 4.18: Real recordings of complex vocalizations from Segde Warbler and Sky Lark [166].

4.5.6 Signal Degradation

As discussed briefly in Section 3.4, a common difficulty with bird song analysis is data is often made up of degraded versions of the signals sent by the birds. The majority of bird recordings are made in the bird's natural habitat. Thick vegetation, like in the canopy of a rainforest, causes signal scattering. High frequency sounds are affected by small obstacles such as leaves while low frequencies are unaffected. This results in a low pass filtered signal arriving at the microphone. Other forms of degradation include discrete echoes, sound tails and cancellations due to the ground effect. Even a bird turning its head while vocalizing can influence the amplitude distribution of the harmonics being picked up by a microphone. Signal degradation was discussed in depth in [96]. A system that segments bird sounds at the end of vocalizations instead of at the end of sound tails would be another useful tool to ornithologists and engineers.

4.6 Discussion

As speech processing techniques motivated work here, it is important to the note the similarities between bird song and speech. Birds and humans both produce vocalizations using a source filter mechanism but straight reapplication of speech processing techniques won't work without some adjustments. Bird vocalizations may sound no more complex than human speech, but recordings are usually subject to more adverse conditions such as contaminant vocalizations and non-homogeneous noise backgrounds [79]. This and the larger bandwidth makes pitch tracking of bird song more complicated than simply applying human PDAs to bird recordings. Pitch is not only important for analysis and synthesis, but is used in measuring bird population similarity. This relies on accurate pitch estimation. Bird frequency range varies dramatically from species to species, and even within syllables in a song repertoire from a single bird. Hence static YIN parameters are not useful in bird recordings. Results presented here have demonstrated that automatically determining the $f_{0_{min}}$ parameter on a segment by segment basis for YIN (performed by YIN-bird) improves pitch estimation. This improvement can in turn improve accuracy on bird species and phrase comparisons, allowing fully automatic batch processing of large numbers of recordings from different species. The synthesized bird calls and ground truth pitch have been shared for research at [114]. This will allow other researchers to compare the performance of PDAs on bird song in a quantitative manner for the first time.

A qualitative description of a range of vocalizations that pose problems for PDAs was also presented in detail and examples shown. Knowledge of when to use YIN-bird, depending on what type of vocalizations are being analyzed, is useful. Pitch estimates of whistles, harmonics with strong $\boldsymbol{F_0}$ and trills should be treated with confidence while pitch tracking on two-voiced and nasal sounds remains problematic.

The value of YIN-bird lies not only in this demonstrated performance improvement. The possibility of fully automatic processing of bird song to extract pitch will allow researchers to

process larger amounts of data, improving certainty in tasks such as species comparison based on pitch measurements.

5

Species delimitation inspired by zoological techniques

Work in this chapter deals with measuring difference between populations in a quantifiable way and identifying the limits of species/subspecies delimitation. Up to now work has taken a signal processing view of how to compare song or calls. Given the work on this topic by zoologists, a more complete exploration of the space necessitates a closer study of how song is used to compare species by the experts, i.e. zoologists. The work of Tobias et al. in [160] gives particular inspiration in this context. Tobias et al. *ibid.*, set out to identify a repeatable method to perform species delimitation. Current practice of species delimitation by zoologists involves manual inspection of spectrograms. Work here sets out to make a move towards automatic performance of such analysis, which removes subjectivity and laboriousness from the acoustic evidence.

The goal of this chapter is to automatically extract the spectral features used *ibid. pg. 11*, to aid ornithologists in their research. The experiments here use a similar dataset to *ibid. pg. 11*. The collection and pre-processing of data is given, followed by a description of expert labeling. Section 5.1.1 gives a brief description of species delimitation system from *ibid. pg. 8*, focusing on acoustic evidence. Section 5.2.2 describes YIN-bird feature extraction and Section 5.2.3 outlines a sine tracking method. Experiments and results are explained in section 5.3.3. In Section 5.4 a comparison between the acoustic feature system used by Tobias et al. and bird difference measure presented in chapter 3 is discussed. The topic of bird population difference was first presented in chapter 2 and chapter 3. The reader is now reminded of work by Harte et al. [59], O'Reilly et al. [116], Meliza et al. [105], and Sangster et al. [137]. Tobias et al. gave a comprehensive background to work on bird species limits in [160]. A brief synopsis of [160]

is given here to give the reader an awareness of species limits and what these experiments are trying to achieve. For a more in depth insight the reader is directed to work *ibid.* Finally conclusions are given in Section 5.5.

5.1 What parts of vocalizations do Zoologists see as important?

Engineering approaches to bird acoustic analysis generally involves using labeled recordings of birds as data, analogous to speaker recognition. The distinction between calls or song is often not important to tasks of species recognition. When zoologists analyze bird populations, they tend to look at acoustic evidence differently. The system *ibid. pg. 11* focuses more attention on mating signals such as songs and displays. The focus was on song rather than calls *ibid. pg. 11* as song tends to function in mate choice and hence reproduction isolation in birds [37]. Song refers to territorial or advertising signals and these are generally identifiable by their complexity or stereotypy in relation to alarm or contact calls [160].

Bird calls have been used extensively in bird classification experiments, [165,57,51,61,38,72, 154,153,79,4,128,28], but calls may not always give clues of species separation or reproductive isolation. Zoologists favor song in their analysis because it is more complex and stereotypical. Song was used exclusively in [138,85,160]. The aim of work in this chapter is to merge techniques used by engineers with approaches applied by zoologists.

5.1.1 Species Delimitaion

Chapter 5 investigates the automation of a taxonomic scoring system presented by Tobias et al. Tobias et al. proposed this simple point-based system where the difference between population pairs was scored according to four degrees of magnitude: minor, medium, major and exceptional. A system where difference is classed like this is quite attractive to engineers and zoologists as it quantifies difference which leads to automatic species/subspecies decision making.

Divergence of different trait types: morphological, voice, and biometrics were summed and compared with irrefutable species[1]. The first step involved quantifying difference in three main types of traits, morphological, voice and biometrics against all four degrees of magnitude. The second step measured ecological and behavioral characters which could only be minor or medium degrees of magnitude. The fourth step assigned taxa to one of four conditions of geographical relationship: allopatry (geographically isolated), broad hybrid zone (geographically distant), narrow hybrid zone (geographically close) and parapatry (geographically adjacent). These were given scores of 0, 1, 2 and 3, respectively. Tobias et al. proposed that the evidence for character differentiation must increase in tandem with geographical separation. In other words, the degree of differentiation required to trigger species status must increase from parapatry through narrow hybrid zones and broad hybrid zones to allopatry. Hence taxa located close together (parapatry)

[1]Irrefutable species defined as species which speciation is not questioned by zoologists [160], pg. 8.

were given a score of 3 (a bias towards species separation) and taxa geographically isolated (allopatry) were given a score of 0. Taxa pairs of subgroups were considered different species if their total score ≥ 7. Subgroups can be two subspecies or species. Upon review of their analysis with this system, a group's taxonomy may change between subspecies and species. Total scores could only include a maximum of two biometric characters (the largest increase, and the largest decrease in effect size), two vocal characters (the largest temporal, and the largest spectral effect size), three plumage characters, and one behavioral or ecological character. Capping of biometric and vocal divergence at two characters allowed the maximum number of characters with the minimum degree of covariance. The system is summarized in Table 5.1 (for a deeper explanation the reader is directed to [160], pg. 8-9). Note a score can reach 7 without analyzing all 5 traits. An example is used to illustrate. The total score when comparing two populations of 'Arremon' was 14. Total score added 2; strongest morphology feature (Tarsus) with a medium difference, 2; strongest temporal feature (duration) with a medium difference, 3; strongest spectral feature (min freq) with major difference, 3; for 1^{st} plumage of major difference (color - midcrown to nape), 2; for 2^{nd} plumage of medium difference (color - supraloral), 2; for 3^{rd} plumage of medium difference (color - supercilium), which summed to 14. Ecology & behavior and geographical relationship were not considered for this comparison *ibid, pg. 7.*

The main goal *ibid.* was to produce a system that is simple and rapid enough to be applied to all birds, quantitative enough to satisfy the need for rigor and repeatability, and transparent enough to allow taxonomic decisions to be traced and evaluated. The goal of this chapter is to automate the extraction of acoustic features using YIN-bird pitch extraction [117] and sine tracking [67], as current practices by zoologists requires manual inspection of spectrograms which is laborious and subjective.

5.1.2 Tobias et al. Experiments

The first set of experiments by Tobias et al. *ibid. pg. 8* used species pairs that are irrefutable to assess phenotypic divergence (divergence of organism's observable characteristics) and to set the limits of Cohen's d (explained in section 5.1.4) which correspond to minor, medium, major and exceptional difference for different traits or features. Various tests were undertaken to investigate capping the number of characters within traits on the mean and variance of total scores, and on the congruence of observer scores for plumage *ibid. pg. 13.* The effect of missing data, song learning and the relationship between divergence in morphology and song across all species were also examined. The threshold for triggering species status was set as the total capped score that resulted in 95% of sympatric taxa (populations who overlap within a geographical regions) being assigned as species (restricting the analysis only to those species pairs with full arrays of data biometric, vocal and visual).

Mean effect sizes were fairly consistent across all biometric and vocal characters, being close to 2.0 in all cases. On the basis of this result the threshold for triggering a medium difference

Table 5.1: Procedure for scoring species pair difference from [160]. If combined totals reach or exceed 7, species status is assigned.

		Magnitude of Difference				**Total**
Trait	**Features**	**Min (1)**	**Med (2)**	**Maj (3)**	**Excep (4)**	**Score**
1. Morphology (biometrics)	Strongest increase & decrease	Effect size: 0.2-2	Effect size: 2-5	Effect size: 5-10	Effect size: >10	1-4
2. Acoustic	Strongest temporal & spectral	Effect size: 0.2-2	Effect size: 2-5	Effect size: 5-10	Effect size: >10	1-4
3. Plumage	3 strongest	Slight diff in wash	Distinct diff in tone	Major diff in hue	Radically diff in coloration	1-4
4. Ecology & behavior	Once off descriptor	E.g. Diff in habit	E.g. Diff in courtship display	N/A	N/A	1-2
5. Geographical relationship	N/A	Broad hybrid zone	Narrow hybrid zone	Parapatry (adjacent)	N/A	0-3
						If sum ≥ 7; species

was set as Cohen's d = 2.0. In addition, the thresholds for major and exceptional differences were arbitrarily set at d = 5.0 and d = 10.0, respectively. Any effect size less than 0.2 was considered biologically irrelevant and hence the threshold for a minor difference was set at d = 0.2 *ibid. pg. 14*. This approach of quantifying difference into minor, medium, major and exceptional is what makes it an attractive scheme for bird song analysis.

When these four thresholds were applied to biometric characters 58.0 ± 5.9% of pairs had a minor difference, 26.0 ± 1.1% had a medium difference, 6.0 ± 1.1% had a major difference and 0.5 ± 1.1% had an exceptional difference (n = 53; 4 character types). A similar proportion of pairs had vocal differences in each of these categories: 63.0 ± 9.0% of pairs had a minor difference, 24.0 ± 6.0% had a medium difference, 3.0 ± 1.0% had a major difference and 1.0 ± 1.0% had an exceptional difference (n = 54; 7 character types). Visual assessment of plumage

characters generated a slightly different category ratio, with higher numbers of medium and major characters, but very few exceptional characters. All but one species pair (98.3%) had at least one minor difference in plumage, 47 pairs (81.0%) had at least one medium difference, 18 pairs (31%) had at least one major difference and only one pair had an exceptional difference.

The second set of experiments used a sample of 23 pairs of subspecies to compare taxonomic recommendations based on Tobias et al.'s approach with those of the BOU guidelines. Unfortunately acoustic evidence was discounted because only a subset of taxa had recordings available and the BOU guidelines do not include information about dealing with acoustic data, making it impossible to standardize acoustic comparisons. These tests revealed that the two approaches produced different taxonomic recommendations. Under the system in Tobias et al. only two pairs of taxa qualify for species status, while 17 pairs of taxa qualify for species status under the BOU guidelines. Under both systems, the proportions of taxa considered species would probably increase with the addition of vocal data. However, as vocal differences between most subspecies are likely to be far smaller than those between sympatric species, Tobias et al. believes that the addition of vocal characters would still produce fewer taxonomic changes under their system than the BOU guidelines *ibid. pg. 15.*

The taxonomic methods outlined in [160] are similar to previous approaches in that decisions are guided by direct reference to known sympatric species. However, the novel addition of a quantitative scoring system allows this calibration to be more accurate and explicit. Results suggest that their approach is workable and relatively consistent. In particular, they confirm that the exclusion of trivial characters, and the capping of character number, limits the accumulation of scores based on minor characters. It also shifts focus to the largest differences between taxa, which are more likely to be associated with reproductive isolation. Testing the criteria against European subspecies limits suggests that the system produces results that align with general standards in Europe concerning biological species. Although these finding may suggest that the thresholds are designed to maintain the taxonomic status quo, it should be kept in mind that vocal data was not included in the analyses, and the addition of this data may trigger species status for some vocally divergent subspecies which were not considered species under previous guidelines *ibid. pg. 16.*

Collar [36] tested an earlier version of the same criteria from [160] against a large suite of Asian taxa, and found that it allowed the separation of 48 new species, of which three have already been supported by independent molecular work [52]. Work in [160] presented a scoring system that can deliver robust and consistent taxonomic change within the conceptual framework of the Biological Species Concept. Another inference from *ibid.* suggests that the methods for capping character number, which were introduced to reduce the difficulty posed by non-independent characters, have the added benefit of reducing uncertainty and maximizing repeatability. Also they reveal an inverse correlation between divergence in plumage and morphology on the one hand and song on the other supporting the idea that there is an evolutionary trade-off between different forms of costly traits. A taxonomy based on morphological

and plumage characters alone will tend to overlook cryptic species whose divergence is biased towards acoustic characters. This highlights the importance of including both morphology and song when assessing species limits in birds *ibid. pg. 16.*

5.1.3 Tobias et al. strengths and limitations of system

It is easy to find fault in any system which delimits species, as species boundaries are inherently fuzzy [70, 94]. No system involving the subdivision of a continuous process can hope to be fully objective [160]. Tobias et al. accept that their system can neither cover every context nor cater for all tastes, but it is hoped that it may help promote accuracy, stability and consistency in avian taxonomy. As many taxonomic decisions rely on distinction between minor and medium characters, slight ambiguities can blur boundaries between species and subspecies *ibid. pg. 17.* Data findings support the hypothesis that reproductive isolation and morphological divergence are not closely correlated, while on the other hand, findings indicate that divergence summed across morphology, plumage, behavior, ecology and acoustic signals is much more likely to be correlated with reproductive isolation *ibid. pg. 17.*

The criteria presented *ibid.* will require further testing against widely accepted species and subspecies limits. They should be viewed as a baseline which can be modified according to new information, including their performance in specific case studies.

Disagreement about what constitutes a species is so pervasive that some authors view current taxonomic assessments of species totals as gross underestimates, whereas others fear that they are overestimates. The work *ibid.* was developed for use with birds, but can easily be adapted for use with other taxa. Taxonomic decision making becomes more uniform and has the power to produce taxonomic changes that are consistent and easily evaluated by independent reviewers. This contrasts with current practice in avian systematics, which generates anything from narrowly divergent allopatric [2] 'species' to highly divergent 'subspecies'. If carefully applied, this system can therefore help to resolve difficult cases with conservation implications and produce a global taxonomy of comparable species units *ibid. pg. 19.* Preliminary tests showed that these criteria result in relatively few changes to avian taxonomy in Europe, yet they might be capable of extensive reassignment of species limits in poorly known tropical regions elsewhere *ibid. pg. 1.*

5.1.4 Acoustic Evidence

A general overview of the system used by Tobias et al. is given in Section 5.1.1. As this thesis concentrates on the field of audio signal processing, only the acoustic evidence measure is investigated here. For more information on morphological, plumage, ecology, behavior and geographical relationships the reader is directed again to *ibid. pg. 11.* Song was used rather than calls, as song tends to function in mate choice and hence reproduction isolation in birds [37]. Song refers to territorial or advertising signals and these are generally identifiable by their

[2] Allopatric populations of related organisms are unable to interbreed because of geographic separation.

complexity or stereotypy in relation to alarm or contact calls. Three temporal and four spectral features were used as part of the vocal evidence. The features extracted from the recordings were as follows:

1. total number of notes
2. duration of song
3. pace (number of notes/duration)
4. maximum frequency
5. minimum frequency
6. bandwidth (max - min)
7. peak frequency (frequency with greatest amplitude)

The spectral features used in [160], pg. 11, were maximum frequency, minimum frequency, bandwidth (maximum minus minimum frequency) and peak frequency (the frequency with the greatest amplitude). These feature definitions are quite broad. Maximum frequency could refer to the maximum F_0 observed during a bout of song, the maximum frequency of the prominent partial or it could mean the maximum frequency of the highest observable partial. For whistles, trills and vocalizations containing a single partial, any of the definitions will produce the same value. The majority of song recordings here contained a strong F_0 partial which could be easily measured. When nasal recordings were observed, maximum frequency was considered the maximum frequency of the prominent partial for experiments here. Recordings at different sampling rates of the same song may exclude high frequency partials resulting in inconsistent maximum frequency measure for the same song. Recording conditions also influence frequency measures, as a song recorded in open conditions may have a wider bandwidth than the same song recorded in dense vegetation where the song is low pass filtered by the environment. The direction of a bird's beak can change during a bout, causing variation in the number of harmonics or partials present. These frequency measurements were taken from the prominent frequency partial, which is fundamental frequency for most cases. In rare cases where fundamental frequency is missing, the prominent partial is analyzed instead. Bandwidth refers to the bandwidth of the pitch and not the bandwidth of all harmonics present, as number of harmonics present in a recording can vary due to recording conditions.

Some practical considerations needed to be taken into account when assessing song divergence. For instance, it was essential that comparisons were made, not only with the corresponding age and sex but also with analogous vocalizations. For example, it would be easy in poorly sampled cases to presume calls of taxon A are comparable to songs of taxon B, or primary song of taxon A to secondary song of taxon B. Doing so would result in spurious estimates of phenotypic divergence. The most diagnostic acoustic traits should be sought when selecting traits

for quantification *ibid. pg. 12.* Testing taxonomic hypotheses with vocal data in birds is also complicated by the issue of learning, which produces complex songs, dialects and repertoires. Taxa in which learning is minimal or absent (i.e. suboscine passerines and most non-passerines) and those in which learning is widespread (i.e. oscine passerines, hummingbirds and parrots) are likely to differ in terms of individual and regional variation *ibid. pg. 12.* Differences in acoustic structure might reflect species status in a suboscine passerine [63], but highlight nothing more than geographical dialects in an oscine passerine. Tobias et al. addressed this issue by restricting measurements to general characters, thereby limiting the influence of character triviality *ibid. pg. 12.*

Once the data was selected for taxa comparisons and features were extracted, the mean and standard deviation was converted to effect size for each feature using Cohen's d statistics. Effect sizes were used because they are more suitable for taxonomic appraisals than P-values, which are highly correlated with sample size [108]. Statistically significant differences can be achieved simply by increasing sample size which may lead to inappropriate taxonomic decisions [120]. Cohen's d statistic is often used for effect size, which combines a measure of the magnitude of a difference with a measure of precision [160], pg. 8. Cohen's d was first published in [34]. Cohen's d was calculated using Eq. 5.1 and Eq. 5.2.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}} \tag{5.1}$$

where $\bar{x}$ *= mean of subgroup 1 and subgroup 2,* s *= standard deviation, and*

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}} \tag{5.2}$$

where n *= number of individuals sampled in species 1 and 2.* [3]

Upon building distributions of effect sizes produced by empirical tests of divergence in undisputed species (see Results Section in [160], pg. 14), vocal differences were scored with an effect size of $0.2 - 2$ as minor, $2 - 5$ as medium, $5 - 10$ as major and > 10 as exceptional. This approach assumes that one can calibrate the significance of effect size differences according to divergence measured across a sample of known species *ibid. pg. 8.* As the acoustic features are heavily correlated, only the strongest temporal feature and spectral feature were used in the acoustic similarity measure *ibid. pg. 8.* The features with largest absolute Cohen's d values were selected as the strongest temporal and spectral feature to use for each pair.

[3] Note *ibid. pg. 12.* the -2 was omitted from the equation but including -2 gave the results presented in supplementary material with [160].

5.2 Automatically Extracting Features

The specific aim in this chapter was to automate the extraction of the frequency features. Acoustic evidence *ibid. pg. 11* was extracted by analyzing spectrograms and using on-screen cursors to measure the features. The introduction of automatic acoustic feature extraction would be of great benefit to zoologists, as manual inspection of spectrograms can be quite tedious and prone to subjectivity.

5.2.1 Temporal Features

The temporal features in Tobias et al. were duration, note number and note pace (duration divided by note number). *Ibid. pg. 12*, these time features were measured from spectrograms with high temporal resolution settings of 1.5 ms. The mean, sample size (number of individual birds measured per population) and standard deviation of these features measured by Tobias et al. in [160] were given in 'IBI_1051_sm_AppendixS2-3.xls' [161]. To automatically measure song duration requires accurate automatic segmentation of bird song recordings. This is attempted in chapter 6. Automatically measuring the number of notes is a more complex task. Notes are generally separated by short pauses that are hard to identify. Sometimes notes run into each other with no pause in between. For these reasons, automatic note numbering was not undertaken for this thesis. Automatic note identification may be possible with complex machine learning and would be an interesting project for future work.

For experiments in this chapter temporal features were calculated manually. Using the labeling information of data described in Section 5.3.1, the same temporal features from [160], pg. 11 were measured on the subset of data collected for experiments in this Chapter. The mean, sample size and standard deviation results were similar but not exactly the same to numbers in [161]. This is understandable as although a similar set of recordings was used there was no way to guarantee the songs chosen were exactly the same. Different sets, with different number of songs, will have different duration and number of notes but the order of measurements was similar which is encouraging. When these features were translated into effect size using Cohen's d statistics the strongest temporal features were not always consistent but for the most part the strongest temporal Cohen's d value fell within the same magnitude of minor, medium, major or exceptional difference.

5.2.2 Feature Extraction using YIN-bird

YIN-bird [117] is a pitch extraction algorithm based on the original YIN [42], but optimized for bird vocalizations. YIN-bird uses spectrogram information to adaptively update the minimum frequency parameter for YIN based on identifying a prominent frequency region on a segment by segment basis within a recording. This prevents octave errors which are frequently observed using original YIN for birdsong. Thus if the fundamental is weak or missing, YIN-bird will track

pitch as the prominent harmonic instead. YIN-bird was introduced in chapter 4.

Bouts of song were segmented from longer wave files in Matlab using the textgrid labeling files described in Section 5.3.2. This meant that $\boldsymbol{F_0}$ or prominent pitch was extracted from target song instances only. Note it was not disclosed in [160] if unvoiced vocalization frequencies were taken into account but for the most part song tended to be melodic and made up purely of voiced instances for the dataset. Therefore the inclusion/exclusion of unvoiced incidences would have little influence on overall mean and variance frequency measures.

The minimum and maximum frequency was selected from YIN-bird's pitch contour output. The peak frequency was calculated by taking the average pitch of YIN-bird's output, i.e. mean of pitch or mean of prominent frequency partial in a minority of cases. The bandwidth per song was calculated by subtracting the minimum frequency from the maximum frequency output from YIN-bird. These four spectral features were collected for every song sample. The mean and standard deviation of these features were then calculated. The feature measurements using YIN-bird.

5.2.3 Feature Extraction using Sine Wave Tracking

The sinusoidal detection algorithm was also employed. This algorithm was introduced by Jančovič et al. in [65], which was used in a number of works on analysis of bird vocalisations and bird species recognition, e.g., [67, 66]. This method performs the detection in the short-time spectral domain. Each peak in the magnitude spectrum of signal frame is considered as a potential sinusoidal component. The decision whether a peak corresponds to a sinusoidal signal or not is based on the maximum likelihood criterion, i.e., the peak is detected as a sinusoid if $p(y|\lambda_s)>p(y|\lambda_n)$, where y is a feature vector consisting of magnitude and phase spectral features extracted around the peak, and λ_n and λ_s are trained models of noise and sinusoidal signals, respectively. Sinusoidal models were trained using simulated sinusoids with a range of linear frequency modulation. The outcome of detection is a set of isolated time-frequency segments, each segment corresponding to a temporal evolution of a sinusoidal component. The initial segmentation was further refined by discarding very short segments and segments of a low energy. In a case of temporal overlap of segments, only the higher energy segment was used.

Note that the code for Sine Wave Tracking described in [67] is not publicly available and the authors were unable to share the code. Hence the sine wave tracking was performed by the original authors on data supplied to them and the results subsequently analyzed.

5.3 Experiments

5.3.1 Data

Tobias et al. *ibid. pg. 11*, used a dataset of vocal measurements from 54 pairs of closely related congeneric (same species) sympatric (populations overlap within a geographical regions)

or parapatric (populations with divergence between different ends of the same geographical regions) species. These species pairs contain representatives from 29 avian families, and were selected in consultation with regional experts *ibid. pg. 11.* As with the bird population difference measure in chapter 3, thresholds needed to be assigned for the system. A detailed list of recordings for these species was given in supplementary material ('IBI_1051_sm_AppendixS2-10.xls' - link available at [161]), along with their source library location. For work in *ibid. pg. 8* recordings were taken from commercially available CDs, xeno-canto.org [166], the Cornell Laboratory of Ornithology Macaulay Library [118], and the British Sound Archive [90]. Recordings were taken within the zone of sympatry where possible. Final samples contained songs from 2 to 10 individuals per species, with 1 to 10 songs per individual. Where there was much intra- and inter-individual variation (as in many oscine species) at least six individuals per species were sampled, and at least six songs per individual. Multiple songs were often analyzed from the same recording. From these recordings, final samples of song were extracted for experiments.

Joseph Tobias was contacted about sharing the final samples, but regrettably they were not available. However, the details contained in the list of recordings allowed the majority of recordings to be acquired independently, from xeno-canto.org [166], the Cornell Laboratory of Ornithology Macaulay Library [118] and the British Sound Archive [90]. Not all of the recordings used in [160] were used here, but all of the recordings used in this chapter were used previously. Of the recordings that were acquired, as many as possible were used. Recordings were manually segmented by Regan Drennan under supervision of the current author. Regan worked as a summer intern with Sigmedia in 2016, bringing with her expertise in bird listening, gained from her zoological studies at Trinity College Dublin and experience from listening in the field.

Species for which data was segmented is given in Table 5.2. Source key: Xeno-canto.org (XC), National Sound Archive, British Museum (NSA) and Library of Natural Sounds, Macaulay, University of Cornell (LNS). The pair number comes from the number given in supplementary material by Tobias et al. [161] and is shown for comparison. Recordings from other species present in [160] but absent from Table 5.2 were also acquired but unfortunately, due to time constraints, could not be annotated.

5.3.2 Expert Labeling of Data using *Praat*

The data introduced in Section 5.3.1 was manually segmented by a zoologist expert in bird song on an internship. As time was limited and this is an extremely time-consuming task, it was not possible to fully segment every file. As some files contained multiple bouts of song, there is no absolute guarantee that the songs selected by Tobias et al. are identical to the songs selected in this work. However it can be assumed with a high degree of confidence that labels are accurate.

The software used for this process was *Praat* [16]. *Praat* is a phonetics software developed by Paul Boersna and David Weenink. It is commonly used by speech processing engineers and linguists when preparing speech data for experiments. It can also be used for bird vocalizations.

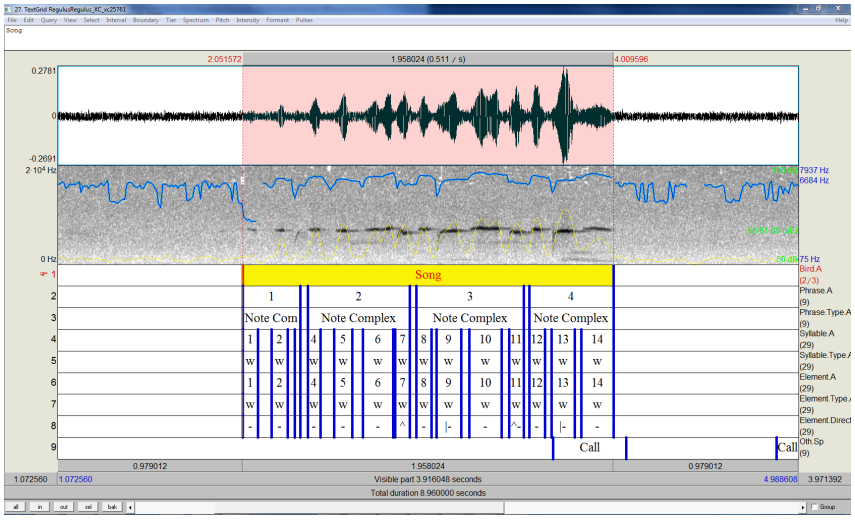
Table 5.2: Species delimitation data. Source key: Xeno-canto.org (XC), National Sound Archive, British Museum (NSA) and Library of Natural Sounds, Macaulay, University of Cornell (LNS).

Pair No.	Common Name	Genus	Species	No. of Files	No. of Songs	Source
5	Chestnut-capped Brush Finch	Arremon	brunneinuchus	4	28	XC
5	White-browed Brush Finch	Arremon	torquatus	3	41	XC
6	Blackish Antbird	Cercomacra	nigrescens	4	19	XC
6	Black Antbird	Cercomacra	serva	4	22	XC
7	Short-toed Treecreeper	Certhia	brachydactyla	4	24	XC
7	Eurasian Treecreeper	Certhia	familiaris	3	13	XC
10	Red-tailed Greenbul	Criniger	calurus	4	24	NSA
10	White-bearded Greenbul	Criniger	ndussumensis	6	47	NSA
12	Hammond's Flycatcher	Empidonax	hammondi	6	50	LNS
12	American Dusky Flycatcher	Empidonax	oberholseri	8	49	LNS
19	Todd's Antwren	Herpsilochmus	stictocephalus	3	9	XC
19	Spot-tailed Antwren	Herpsilochmus	sticturus	2	6	XC
41	Downy Woodpecker	Picoides	pubescens	6	6	LNS
41	Hairy Woodpecker	Picoides	villosus	3	22	LNS
44	Common Firecrest	Regulus	ignicapillus	3	35	XC
44	Goldcrest	Regulus	regulus	3	27	XC
45	Louisiana Waterthrush	Seriurus	motacilla	5	58	LNS
45	Northern Waterthrush	Seiurus	noveboracensis	9	50	LNS
53	Mouse-coloured Antshrike	thamnophilus	murinus	3	5	XC
53	Plain-winged Antshrike	thamnophilus	schistaceus	3	11	XC
54	Bendire's Thrasher	Toxostoma	bendirei	3	15	LNS
54	Curve-billed Thrasher	Toxostoma	curviroste	5	15	LNS
55	Western Kingbird	Tyrannus	vericalis	1	50	LNS
55	Cassin's Kingbird	Tyrannus	vociferans	5	51	LNS
57	Golden Winged Warbler	Vermivora	chrysoptera	8	50	LNS
57	Blue-winged Warbler	Vermivora	pinus	10	50	LNS
58	Red-eyed Vireo	Vireo	olivaceus	9	100	LNS
58	Philadelphia Vireo	Vireo	philadelphicus	10	100	LNS

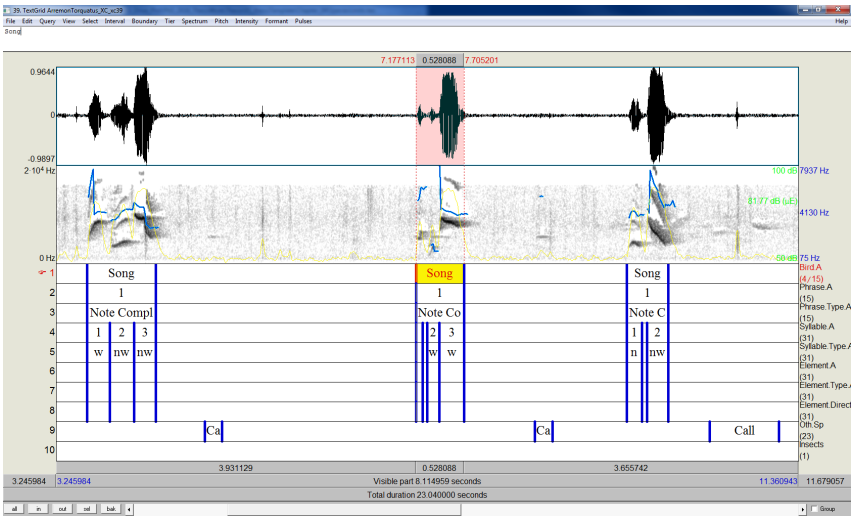
Praat was used for visual spectrogram inspection of bird song recordings to annotate phrase identity and time boundaries of phrases by Kaewtip et al. in [71]. The steps involved in labeling bird recordings were as follows:

1. Open wave file of bird recording in *Praat*.
2. View time and frequency domain plots.
3. Create tiers for each level of segmentation. Levels of segmentation were as follow:
   (a) Song or call
   (b) Phrase number
   (c) Phrase type
   (d) Syllable number
   (e) Syllable type
   (f) Element number
   (g) Element type
   (h) Element pitch direction
4. Manually select start and end times of vocal events at each level and add text label. This was done by a combination of visual inspection of spectrogram and aural analysis of recordings.
5. If multiple birds are present in recording add more tiers.
6. Save the information as a textgrid file. Each wave file has its own corresponding textgrid annotation file.

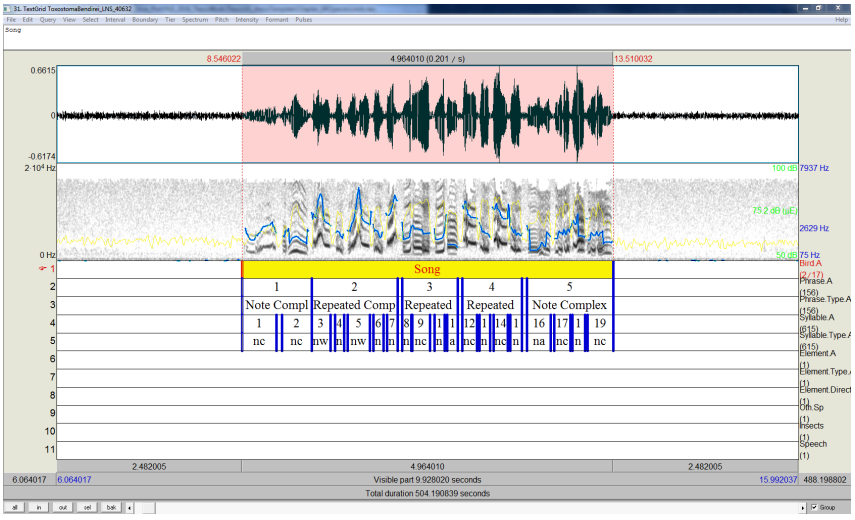
The majority of files contained vocalizations of target birds in the foreground, with non-target birds faintly observed in the background. Screenshots of *Praat* can be seen in Figure 5.1. Figure 5.1a shows a single bout of Regulus regulus, with song, phrase, syllable and element levels fully annotated. Note a faint call from a non-target bird is observed and labeled on the "Other Species" tier. Figure 5.1b shows song of Arremon torquatus with zoom settings set to view three bouts. Note that at this zoom settings some of the longer text between boundaries disappears from view, e.g. at 7.177 seconds, during the second song, the "1" is hidden from the 'Syllable.A' tier, as are all element level annotations. Figure 5.1c shows a complex song of Toxostoma bendire. Towards the end of the internship, in order to save time, only song, phrase and syllable level annotations were made due to the fact work in this thesis concentrated on song and syllable level analysis. While not used here the labeling of phrases and elements will be of use to future research.



(a) Regulus regulus *Praat* screenshot



(b) Arremon torquatus *Praat* screenshot



(c) Toxostoma bendirei *Praat* screenshot

Figure 5.1: Screenshots of *Praat* showing time (top) and frequency (middle) domain information along with thorough labeling of boundaries (bottom).

The main tier of interest for this chapter is the top tier 'Bird.A', which contains song level annotation. Labels for this tier were either song, incomplete song, false start or call. Song is described as a distinct series of phrases and is generally long or continuous. Calls are usually short, monotonous and simple. Bouts of song were contained within start and end boundaries on the 'Bird.A' tier. These bouts were extracted from files for feature extraction. Information outside of these boundaries, with no text, was considered background noise, silence or non-target vocalization and was not used for feature extraction. Target non-song was also excluded from feature extraction. The next tier was entitled 'Phrase.A', which encompassed the phrase boundaries. Each phrase was labeled by its number of appearance, 1, 2, 3,..., n, where n was the number of phrases in the song. Tier 'Phrase.Type.A' corresponds to the previous tier with the same boundaries but instead of numbering, phrases are described as either note complex, trill complex, repeated notes, single note, flourish, trill - new addition or repeated complexes. Phrases are considered distinct sections of song that contain a series of units, often repeated. Phrases may be repeated themselves, or be delivered in a set sequence with other phrases. A song may contain multiple phrases, equally however, a song may be composed of only one phrase and this is conditioned on subjectivity. The description note complex is a broad term. At the phrase level, it includes a distinct set of syllables that are not the same, either by syllable type and/or frequency/duration. At the syllable level (this only really occurred in fast complex warbled songs) it refers to a syllable which consists of elements that differ in type, e.g. a single syllable that consists of a whistle, a burr, and a nasal note very close together. Repeated notes were most commonly the same note repeated, but also included the same notes repeated at slightly different speeds/frequencies, or similar notes at the same speed and frequency.

One of the acoustic features used by Tobias et al. was number of notes. This feature was calculated from annotations at the syllable level. A syllable is defined as a unit within a phrase, usually one sound, analogous to syllables in speech. Syllables can be simply one element or made up of many elements (complex syllable). Syllable selection can again be subjective. What constitutes a syllable, or note, according to means in Tobias et al. [160], was quite variable. Syllable selection was based partially by ear and partially by visual inspection of notes on the spectrogram. Final selection was somewhat influenced by numbers in Tobias et al. supplementary material [161]. Tier 'Syllable.A' consisted of start and end boundaries for syllables and syllable number label. Tier 'Syllable.Type.A' described each syllable as either a whistle (w), trill (t), note complex (nc), nasal note (na), nasal whistle (nw), noisy (ns), noisy whistle (nsw), burry (br), buzzy (bz) or chirp (c). When polyphony or the two-voiced phenomenon was observed, "p" was added to the description. The next level of segmentation was the element level. Bird song units were described and defined in Section 2.3 of chapter 2. Often elements and syllables were equal, with the term note used for either element or syllable. Although the hierarchy generally is element or note, syllable, phrase, song, the feature 'number of notes' is actually the number of syllables for work here. The tier 'Element.A' accommodates the start and end boundaries along with the element number while the parallel tier 'Element.Type.A' describes the element akin

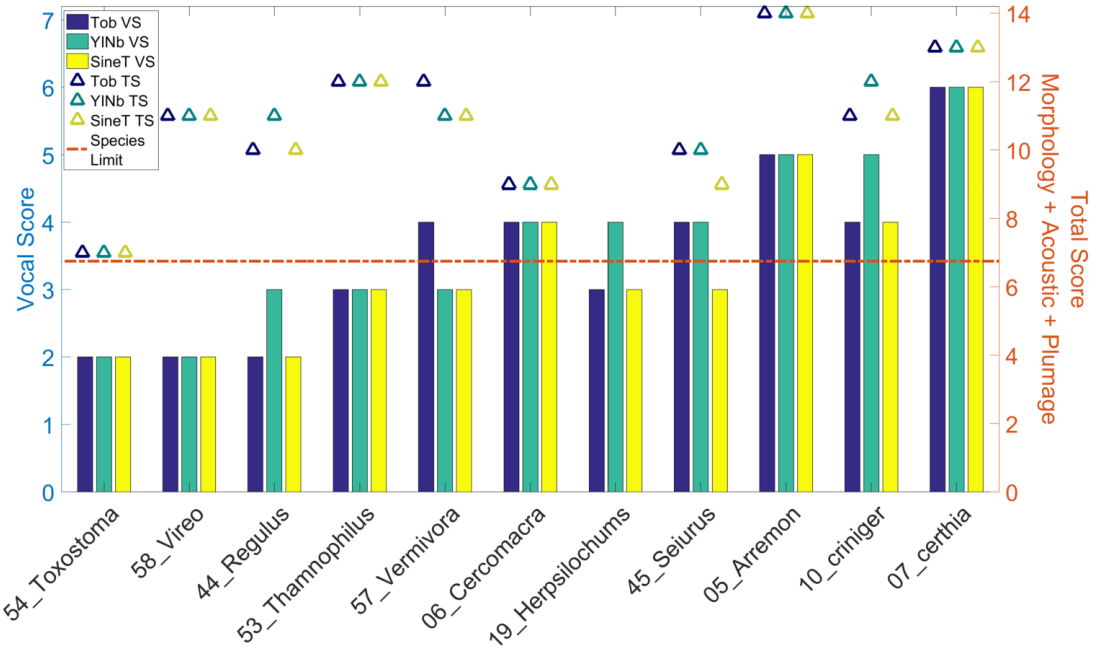


Figure 5.2: Bar chart shows vocal scores from Tob (navy), using features extracted by YINb (green), using features extracted by SineT (yellow) measured on the left y-axis. Total scores from [160, 161], which sums vocal, plumage and morphology scores, are plotted with blue triangular markers using the y-axis to the right. Total scores using YINb and SineT are plotted with green and yellow markers respectively.

to the description of syllables. Finally tier 'Element.Direction' characterizes the pitch contour shape or direction with the key: straight (−), upslurr (/), downslurr (\), overslurr (∧), underslurr (v), sigmoidal (S), straight wavy (∼), chirp burst (|). Other tiers contained boundaries for non-target bird events labeled as call, song, mixed chorus (sections in which more than one bird is singing in the background) or background chorus (when a mixed chorus of other birds are singing for the entirety of the track). Insects and recordist speech were also identified in the annotation file.

5.3.3 Results

Bouts of song were segmented from longer wave files in Matlab using the textgrid labeling files. This meant that between 15 and 200 songs were analyzed per species pair, with an average of 70 songs. Only the number of individuals was given in the original work. YIN-bird and sine tracking were both applied and the spectral features extracted from both as described in Sections 5.2.2 and 5.2.3. The extracted spectral features were combined with temporal acoustic features given in supplementary material of [160, 161] to yield the overall vocal score. A summary of how well the resulting vocal scores calculated by automatic means matched the score manually

calculated by Tobias et al. is presented in Figure 5.2.

Figure 5.2 shows vocal scores (VS) from [160, 161] (label Tob, navy bars), vocal scores calculated using features extracted by YIN-bird (label YINb, green bars), and vocal scores calculated using features extracted by Sine Tracking (label SineT, yellow bars) measured on the left y-axis. The x-axis contains each pair's genus name, preceded by its reference number (1-58) from [160, 161]. This allows direct reference back to the original work. Biometric and plumage scores from supplementary material were combined with vocal scores to aggreagte total scores (TS). TS using Tob are plotted with blue triangular markers using the y-axis to the right. TS using YINb and SineT are plotted with green and yellow markers respectively. TS (morphology + acoustic + plumage evidence) are included in Figure 5.2 to show how vocal evidence influences the final difference score using the system from [160]. In [160, 161], there was no biometric data available for genus 19 'Herpsilochums' hence it does not have a TS plotted in Figure 5.2. Species status is maintained if the TS is ≥ 7. All these pairs maintain species status using all methods which suggests automatic methods adequately agree with Tob VS difference. Circumstances where other evidence sums to 4, VS of 3 will maintain species status, while VS of 2 means pairs are considered subspecies. Here a variation in VS by $\pm$ 1 is critical to a pairs' evaluation.

6 of 11 pairs had the same vocal score using all 3 methods of feature extraction. When using YINb, 7 of 11 pairs had the same score as Tob. For SineT, 9 of 11 pairs obtained the same score as Tobias et al. Vocal scores from all pairs were within $\pm$1 between feature extraction methods (vocal score is given in the right hand column of Table 5.3). 3 of 11 pairs had the same strongest feature using Tob, YINb and SineT methods. For YINb, 5 of 11 pairs had the same strongest spectral feature as Tob. For SineT, 3 of 11 pairs had the same strongest spectral feature as Tob. From the point of view of calculating TS for species status, degree of difference is more important than which spectral feature is used, nonetheless the strongest feature to use should ideally not change between feature extraction method.

Scores are examined in more detail in Table 5.3. The methods of Tobias et al. (Tob), YINb by O'Reilly (O'R), and SineT by Jančovič et al. used to calculate temporal and spectral features are given. The first column, 'Method', states which experiment the figures from that row belong to. The next group of columns, 'Details', contains the pair number from [160, 161], the genus and the species. The adjacent columns, 'Mean', gives the average feature value per taxa population for 'Duration' (Dur), 'Number of Notes' (Note), 'Pace' (number of notes/duration), 'Minimum frequency' (Min), 'Maximum frequency' (Max), 'Peak frequency' (Peak), and 'Bandwidth' (BW). The next set of columns, 'Sample Size', contains a column for, 'Bird Count' (number of different birds recorded for a given taxa), and 'Song Count' (number of songs analyzed per taxa). The next group of cells, 'Standard Deviation', give the standard deviation of these features. Cohen's d statistics were calculated on a pair by pair basis using the method described in Section 5.1.4. The Cohen's d values for temporal features, duration, note and pace, are highlighted in blue with the strongest temporal feature in bold blue. The Cohen's d values for spectral features, maximum, minimum, peak and bandwidth frequencies, are highlighted in

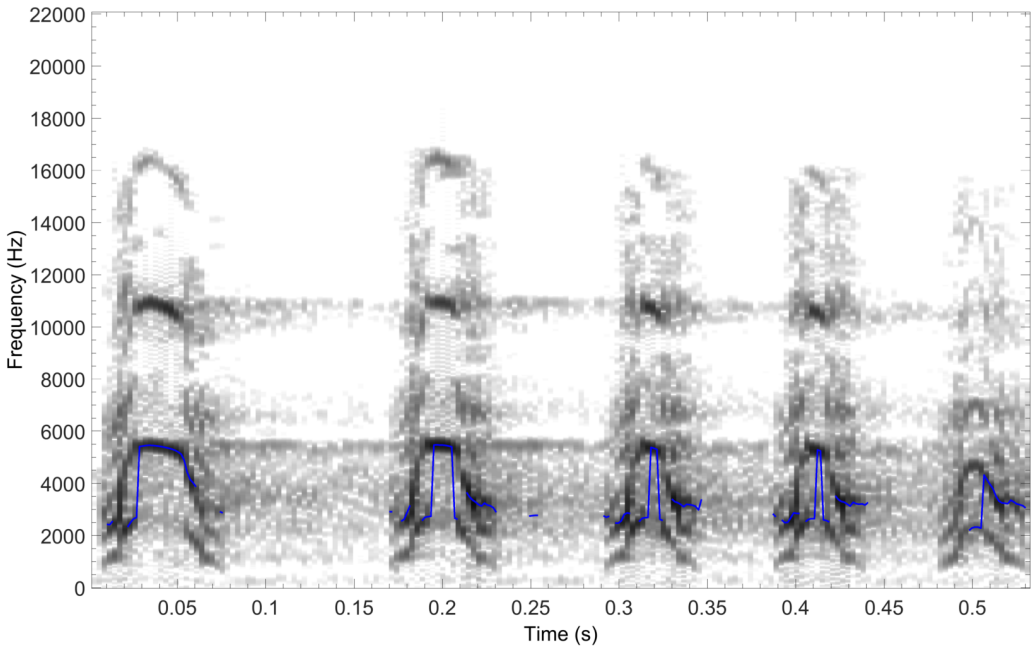
green with the strongest spectral feature in bold green in Table 5.3.

For the most part all experiments agreed on the strongest temporal and spectral features for each pair. The Cohen's d measure for these features also generally fell within the same difference interval of minor, medium, major or exceptional. Data from 14 pairs of taxa was labeled. Of these 14 pairs, 11 pairs contained song that F_0 could be accurately tracked using YIN-bird. YIN-bird could not automatically extract the spectral features for Picoides pubescens and Picoides villosus (pair 41) as their song contained nasally song with the strongest partial varying between the fundamental and 2nd partial (see Figure 5.3a). Tyrannus verticalis and Tyrannus vociferans (pair 55) song have the same nasally characteristic as the Picodies pair, with complex syllables present in the song also (see Figure 5.3b), which adds to the difficulty of automatically extracting their spectral features. Empidonax (pair 12) also has vocalizations that YIN-bird could not accurately calculate. For the other 11 taxa pairs, YIN-bird was able to be used to accurately extract the acoustic features. Sine tracking was able to extract features from all 14 pairs but the three aforementioned pairs required manual selection of which sine track to use.

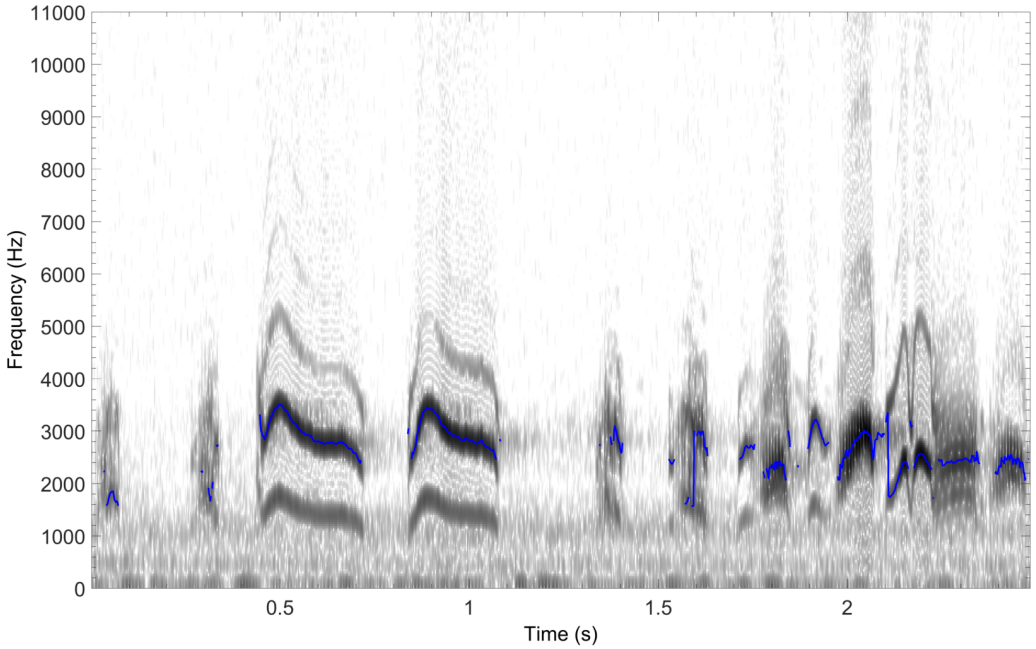
Pair 5, 'Arremon' obtained vocal difference scores which agree across all three feature extraction methods. The strongest feature is minimum frequency for all methods. The mean and standard deviation values of features, for both Arremon populations, occur within a similar range which demonstrates that the use of automatic feature extraction methods produces the same finding as manual inspection analysis.

Pair 44, 'Regulus' presents feature values that do not agree across methods. Mean minimum frequency values for Regulus ignicapillus were 4.30 (Tob), 6.38 (YINb) and 6.43 (SineT) kHz. The mean minimum frequency for Regulus regulus were 4.15, 4.91 and 5.01 kHz. Both the automatic methods output a higher minimum frequency value than Tobias et al. An example of song from Regulus ignicapillus in shown in Figure 5.4. The song of Regulus contains rapid rising pitch modulations which appear stretched on a spectrogram even at high resolution. The true pitch may minimize at 5 or 6 kHz but due to its slope, appears as a dark blur with edges at 4.30 or 4.15 kHz on a spectrogram which may explain the discrepancy in values. YINb and SineT pitch are superimposed along with lines describing the population mean minimum frequency found by Tob, YINb and SineT methods. The two automatic methods give a higher mean minimum frequency than observed when manually inspecting the spectrogram. Both interpretations are fair, but from a signal processing point of view the true pitch calculated is more objective than how song appears on a spectrogram. Standardization of feature extraction requirements are necessary for future use of this system.

The feature with the strongest Cohen's d values also disagrees. Minor difference based on maximum frequency was found by Tob, medium difference using peak frequency was found using YINb and minor difference using minimum frequency was found by SineT. These inconsistencies are most likely due to the difficulty of tracking of syllables present in some examples of Regulus ignicapillus.



(a) Example of Picoides villosus song with nasally quality, which proves difficult for PDAs to extract acoustic features.



(b) Example of Tyrannus vociferans song with nasally quality and complex syllables which prove difficult for PDAs to extract acoustic features.

Figure 5.3: Example of song where accurate acoustic feature extraction could not be performed due to their complexity.

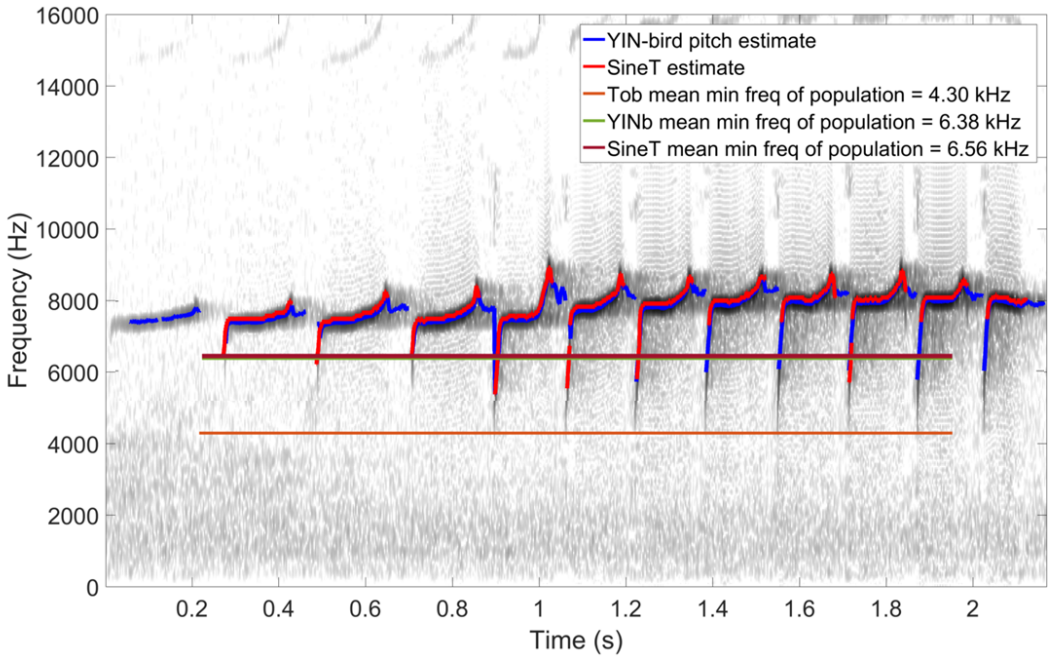


Figure 5.4: Song of Regulus ignicapillus with pitch estimate from YINb (blue) and SineT (red) superimposed. The mean minimum frequency, of all ignicapillus song examples, using the methods of Tob (orange), YINb (green) and SineT (red).

Table 5.3: Results of vocal difference using different feature extraction methods. The Cohen's d values for temporal features, duration, note and pace, are highlighted in blue with the strongest temporal feature in bold blue. The Cohen's d values for spectral features, maximum, minimum, peak and bandwidth frequencies, are highlighted in green with the strongest spectral feature in bold green. The strongest temporal and spectral values are assigned a magnitude of difference. The maginitude of difference is used to compute the vocal score given in the last column, e.g. one medium difference plus one major difference equates to a vocal score of 5; 1x2 + 1x3 = 5.

	Details			Mean							Sample Size		Standard Deviation							Cohen's d							Min	Med	Maj	Exc	
Met-hod	Pair	Genus	Species	Dur (s)	Note	Pace	Max kHz	Min	Peak	BW	Bird Count	Song Count	Dur	Note	Pace	Max	Min	Peak	BW	Dur	Note	Pace	Max	Min	Peak	BW	x1 0.2−2	x2 2−5	x3 5−10	x4 > 10	Vocal Score
Tob	5	Arremon	brunneinuchus	1.48	6.19	4.21	10.84	6.20	9.48	4.64	4	?	0.26	1.21	0.73	0.37	0.70	0.44	0.42	**3.55**	3.20	-0.42	-0.78	**5.65**	4.18	-4.29	0	1	1	0	5
Tob		Arremon	torquatus	0.65	2.89	4.45	11.39	2.98	6.86	8.40	3	?	0.19	0.67	0.23	1.01	0.28	0.83	1.29												
YINb	5	Arremon	brunneinuchus	1.56	6.11	3.94	10.35	7.31	9.08	3.05	4	28	0.17	1.69	0.99	0.39	0.46	0.40	0.41	**4.49**	2.31	-0.09	0.86	**6.74**	5.86	-3.04	0	1	1	0	5
YINb		Arremon	torquatus	0.73	2.88	4.02	9.64	3.55	6.49	6.10	3	41	0.19	0.78	0.68	1.21	0.68	0.51	1.51												
SineT	5	Arremon	brunneinuchus	1.56	6.11	3.94	10.29	7.61	9.20	2.68	4	28	0.17	1.69	0.99	0.43	0.55	0.45	0.45	**4.49**	2.31	-0.09	1.50	**5.41**	2.55	-2.94	0	1	1	0	5
SineT		Arremon	torquatus	0.73	2.88	4.02	9.17	3.58	7.10	5.59	3	41	0.19	0.78	0.68	1.06	0.96	1.19	1.46												
Tob	6	Cercomacra	nigrescens	1.01	2.00	2.02	3.22	1.34	2.66	1.88	4	?	0.14	0.00	0.28	0.17	0.19	0.13	0.23	-3.97	**-4.56**	-3.04	0.26	0.36	**1.89**	-0.09	0	2	0	0	4
Tob		Cercomacra	serva	2.38	7.23	3.05	3.17	1.28	2.31	1.90	6	?	0.42	1.45	0.37	0.17	0.17	0.21	0.22												
YINb	6	Cercomacra	nigrescens	1.08	1.89	1.77	2.86	1.80	2.42	1.05	4	19	0.12	0.32	0.29	0.11	0.13	0.13	0.08	-2.78	-3.39	**-4.17**	-0.87	1.90	0.53	**-2.68**	0	2	0	0	4
YINb		Cercomacra	serva	2.38	7.73	3.24	3.03	1.46	2.33	1.56	4	22	0.66	2.41	0.41	0.25	0.22	0.22	0.26												
SineT	6	Cercomacra	nigrescens	1.08	1.89	1.77	2.93	1.84	2.72	1.09	4	19	0.12	0.32	0.29	0.12	0.19	0.23	0.13	-2.78	-3.39	**-4.17**	-0.66	1.86	0.63	**-3.02**	0	2	0	0	4
SineT		Cercomacra	serva	2.38	7.73	3.24	3.04	1.47	2.59	1.56	4	22	0.66	2.41	0.41	0.21	0.20	0.18	0.18												
Tob	7	Certhia	brachydactyla	1.05	5.75	5.54	7.67	4.57	5.41	3.10	4	?	0.18	0.96	0.99	0.25	0.24	0.62	0.36	**-5.00**	-3.02	-0.85	-1.39	**7.43**	-0.88	-4.50	0	0	2	0	6
Tob		Certhia	familiaris	2.42	16.08	6.51	8.13	3.06	5.95	5.07	3	?	0.37	5.28	1.33	0.42	0.13	0.59	0.54												
YINb	7	Certhia	brachydactyla	1.16	5.67	4.87	7.32	4.96	5.91	2.36	4	24	0.17	0.96	0.37	0.23	0.22	0.19	0.20	**-7.18**	-5.81	-3.60	-1.65	5.84	-0.46	**-6.45**	0	0	2	0	6
YINb		Certhia	familiaris	2.79	18.85	6.70	7.65	3.69	5.99	3.96	3	13	0.29	3.39	0.66	0.14	0.23	0.13	0.31												
SineT	7	Certhia	brachydactyla	1.16	5.67	4.87	7.45	4.99	5.46	2.45	4	24	0.17	0.96	0.37	0.25	0.26	0.50	0.21	**-7.18**	-5.81	-3.60	-3.21	6.59	-0.93	**-7.07**	0	0	2	0	6
SineT		Certhia	familiaris	2.79	18.85	6.70	8.31	3.32	5.97	4.99	3	13	0.29	3.39	0.66	0.30	0.24	0.61	0.51												
Tob	10	Criniger	calurus	1.11	4.21	3.98	3.77	1.69	2.43	2.08	6	?	0.34	0.75	0.77	0.53	0.11	0.15	0.58	0.94	**1.36**	0.20	1.90	3.20	**6.23**	0.98	1	0	1	0	4
Tob		Criniger	ndussumensis	0.84	3.19	3.86	3.06	1.38	1.71	1.68	8	?	0.25	0.75	0.36	0.22	0.09	0.08	0.22												
YINb	10	Criniger	calurus	1.30	4.58	3.54	3.83	1.85	2.48	1.98	4	24	0.23	0.83	0.29	0.69	0.06	0.10	0.69	**3.12**	3.08	-0.45	1.81	3.24	**7.14**	1.09	0	1	1	0	5
YINb		Criniger	ndussumensis	0.82	2.99	3.67	2.97	1.52	1.84	1.45	6	47	0.08	0.15	0.31	0.28	0.12	0.08	0.3												
SineT	10	Criniger	calurus	1.30	4.58	3.54	3.50	1.79	2.92	1.71	4	24	0.23	0.83	0.29	0.30	0.09	0.16	0.29	**3.12**	3.08	-0.45	2.28	**3.96**	2.22	1.08	0	2	0	0	4
SineT		Criniger	ndussumensis	0.82	2.99	3.67	2.86	1.47	2.39	1.39	5	37	0.08	0.15	0.31	0.26	0.07	0.28	0.30												
Tob	12	Empidonax	hammondi	1.89	6.22	3.99	7.83	1.97	5.22	5.86	7	?	0.68	0.66	0.89	0.41	0.12	0.78	0.40	**0.41**	-0.02	-0.25	**2.43**	-0.40	1.55	2.30	1	1	0	0	3
Tob		Empidonax	oberholseri	1.64	6.24	4.17	6.71	2.02	4.16	4.69	8	?	0.54	1.43	0.50	0.49	0.15	0.59	0.58												
YINb	12	Empidonax	hammondi	2.16	6.00	2.84					6	50	0.36	0.00	0.44					**0.93**	-0.44	-0.81									
YINb		Empidonax	oberholseri	1.79	7.39	4.21					8	49	0.43	4.10	2.19																
SineT	12	Empidonax	hammondi	2.16	6.00	2.84	7.48	2.56	6.13	4.93	6	50	0.36	0.00	0.44	0.54	0.22	0.37	0.60	**0.93**	-0.44	-0.81	2.00	-1.17	**3.41**	2.24	1	1	0	0	3
SineT		Empidonax	oberholseri	1.79	7.39	4.21	6.43	2.90	4.86	3.53	8	50	0.43	4.10	2.19	0.51	0.34	0.37	0.64												
Tob	19	Herpsilochmus	stictocephalus	2.05	14.96	7.18	2.95	1.08	2.27	1.87	4	?	0.36	5.18	1.14	0.33	0.24	0.12	0.56	-0.06	0.04	**0.27**	-0.92	0.97	**-2.21**	-0.99	1	1	0	0	3
Tob		Herpsilochmus	sticturus	2.07	14.72	6.85	3.29	0.88	2.56	2.41	3	?	0.63	7.25	1.33	0.42	0.13	0.15	0.52												
YINb	19	Herpsilochmus	stictocephalus	1.89	12.56	6.63	3.52	1.58	2.29	1.95	3	9	0.03	0.88	0.41	2.13	0.09	0.11	2.14	-1.31	**-2.37**	-2.45	0.11	-1.12	**-2.42**	0.20	0	2	0	0	4
YINb		Herpsilochmus	sticturus	2.43	20.17	8.45	3.33	1.73	2.57	1.60	2	6	0.70	5.42	1.14	0.23	0.20	0.13	0.41												
SineT	19	Herpsilochmus	stictocephalus	1.89	12.56	6.63	2.86	1.41	2.49	1.45	3	9	0.03	0.88	0.41	0.15	0.28	0.22	0.28	-1.31	**-2.37**	-2.45	**-1.80**	-0.26	-0.42	-1.59	1	1	0	0	3
SineT		Herpsilochmus	sticturus	2.43	20.17	8.45	3.92	1.48	2.57	2.44	2	6	0.70	5.42	1.14	1.00	0.24	0.09	1.00												
Tob	41	Picoides	pubescens	1.38	13.67	10.09	7.23	2.22	6.65	5.01	6	?	0.38	2.94	1.61	0.34	0.57	0.27	0.52	**-0.90**	-0.85	0.08	-1.31	1.76	**4.86**	-1.83	1	1	0	0	3

Table 5.3 - Continued on next page.

Table 5.3 - Continued from previous page.

Tob		Picoides	villosus	4.17	27.35	9.83	8.83	1.36	3.97	7.47	5	?	4.62	23.89	4.53	1.79	0.37	0.77	1.93												
YINb	41	Picoides	pubescens	1.40	13.50	9.81					6	6	0.42	3.27	1.67					**-1.13**	-0.65	0.19								
YINb		Picoides	villosus	3.76	32.86	9.10					3	22	3.85	55.22	6.72															
SineT	41	Picoides	pubescens	1.40	13.50	9.81	7.09	5.12	6.70	1.97	6	6	0.42	3.27	1.67	0.31	0.76	0.35	0.69	**-1.13**	-0.65	0.19	0.91	4.36	**3.74**	-3.20	1	1	0	0	3
SineT		Picoides	villosus	3.76	32.86	9.10	6.53	2.09	4.30	4.44	3	22	3.85	55.22	6.72	1.04	0.51	1.06	0.95												
Tob	44	Regulus	ignicapillus	2.51	23.20	9.21	8.85	4.30	7.21	4.55	3	?	0.53	5.91	0.52	0.45	1.23	0.30	1.21	-0.18	0.20	**1.60**	**-0.62**	0.12	-0.45	-0.40	2	0	0	0	2
Tob		Regulus	regulus	2.59	22.18	8.49	9.24	4.15	7.31	5.09	3	?	0.37	3.95	0.37	0.76	1.30	0.13	1.45												
YINb	44	Regulus	ignicapillus	2.21	14.46	6.53	8.34	6.38	7.45	1.96	3	35	0.49	4.08	1.04	0.40	0.41	0.19	0.73	-1.01	**-1.73**	-1.66	0.94	1.96	**2.18**	-1.29	1	1	0	0	3
YINb		Regulus	regulus	2.90	22.85	8.03	8.00	4.91	7.00	3.09	3	27	0.83	5.52	0.73	0.32	0.98	0.18	1.00												
SineT	44	Regulus	ignicapillus	2.21	14.46	6.53	8.45	6.43	7.54	2.02	3	35	0.49	4.08	1.04	0.40	0.48	0.29	0.76	-1.01	**-1.73**	-1.66	1.40	**1.96**	1.12	-1.21	2	0	0	0	2
SineT		Regulus	regulus	2.90	22.85	8.03	7.97	5.01	6.90	2.96	3	27	0.83	5.52	0.73	0.27	0.90	0.74	0.79												
Tob	45	Seiurus	motacilla	2.07	9.21	4.37	8.17	2.16	4.80	6.00	5	?	0.20	2.11	0.77	0.85	0.16	0.72	0.91	1.24	-1.92	**-3.24**	0.20	**2.50**	0.45	-0.35	0	2	0	0	4
Tob		Seiurus	noveboracensis	1.75	13.89	7.97	8.03	1.78	4.50	6.25	9	?	0.28	2.59	1.25	0.52	0.15	0.64	0.59												
YINb	45	Seiurus	motacilla	2.02	8.54	4.15	7.58	2.93	4.93	4.66	5	58	0.23	2.60	0.86	0.80	0.51	0.47	1.12	0.51	-1.21	**-2.07**	0.39	**2.53**	1.53	-0.69	0	2	0	0	4
YINb		Seiurus	noveboracensis	1.90	10.84	5.75	7.33	2.11	4.47	5.22	9	50	0.24	1.45	0.72	0.55	0.16	0.16	0.59												
SineT	45	Seiurus	motacilla	2.02	8.54	4.15	7.27	2.50	4.91	4.77	5	58	0.23	2.60	0.86	0.59	0.68	0.68	0.95	0.51	-1.21	**-2.07**	0.03	**1.41**	0.42	-0.94	1	1	0	0	3
SineT		Seiurus	noveboracensis	1.90	10.84	5.75	7.28	1.87	4.65	5.41	9	50	0.24	1.45	0.72	0.50	0.26	0.59	0.50												
Tob	53	Thamnophilus	murinus	2.10	5.25	2.49	2.37	1.07	1.84	1.30	4	?	0.18	0.96	0.29	0.19	0.17	0.10	0.10	-0.43	-4.39	**-4.69**	0.85	0.00	-0.21	**1.18**	1	1	0	0	3
Tob		Thamnophilus	schistaceus	2.18	9.67	4.45	2.25	1.07	1.86	1.18	4	?	0.21	1.05	0.52	0.05	0.14	0.12	0.10												
YINb	53	Thamnophilus	murinus	2.21	5.80	2.59	1.91	1.43	1.74	0.48	3	5	0.25	1.48	0.40	0.10	0.05	0.05	0.07	-0.06	-2.74	**-3.15**	-0.14	**-0.96**	-0.40	0.42	1	1	0	0	3
YINb		Thamnophilus	schistaceus	2.23	9.09	4.11	1.93	1.48	1.77	0.45	3	11	0.18	0.83	0.56	0.09	0.05	0.09	0.09												
SineT	53	Thamnophilus	murinus	2.21	5.80	2.59	2.11	1.37	1.75	0.74	3	5	0.25	1.48	0.40	0.18	0.05	0.09	0.19	-0.06	-2.74	**-3.15**	**0.68**	0.03	-0.49	0.61	1	1	0	0	3
SineT		Thamnophilus	schistaceus	2.23	9.09	4.11	2.01	1.37	1.80	0.65	3	11	0.18	0.83	0.56	0.09	0.12	0.10	0.11												
Tob	54	Toxostoma	bendirei	5.45	43.61	8.03	5.64	1.14	2.56	4.50	5	?	2.48	21.95	2.79	1.23	0.21	0.39	1.26	0.51	1.20	**1.26**	-0.18	0.30	**0.77**	-0.23	2	0	0	0	2
Tob		Toxostoma	curvirostre	4.44	23.94	5.49	5.82	1.08	2.32	4.74	5	?	1.25	7.69	0.60	0.70	0.20	0.17	0.84												
YINb	54	Toxostoma	bendirei	10.39	38.20	3.70	6.13	1.26	2.66	4.87	3	15	8.07	29.77	0.55	0.70	0.33	0.21	0.82	0.82	0.66	**-0.84**	0.16	0.11	**0.83**	0.13	2	0	0	0	2
YINb		Toxostoma	curvirostre	6.12	25.27	4.19	5.89	1.23	2.46	4.66	5	15	2.94	11.16	0.60	1.83	0.34	0.26	1.92												
SineT	54	Toxostoma	bendirei	10.39	38.20	3.70	5.12	1.26	3.66	3.86	3	15	8.07	29.77	0.55	0.70	0.18	0.65	0.69	0.82	0.66	**-0.84**	0.25	0.68	**1.76**	0.14	2	0	0	0	2
SineT		Toxostoma	curvirostre	6.12	25.27	4.19	4.89	1.17	2.62	3.73	5	15	2.94	11.16	0.60	0.99	0.12	0.57	1.03												
Tob	55	Tyrannus	verticalis	1.83	17.08	9.84	5.78	1.35	3.00	4.43	3	?	0.35	4.47	3.52	0.37	0.33	0.75	0.69	-0.71	1.50	**2.68**	**5.71**	1.93	0.15	3.36	0	1	1	0	5
Tob		Tyrannus	vociferans	2.38	11.25	4.92	4.16	1.01	2.95	3.15	8	?	0.86	3.69	0.89	0.25	0.08	0.20	0.22												
YINb	55	Tyrannus	verticalis	2.28	14.36	6.36					1	50	0.62	6.48	1.73					-0.45	0.86	**1.97**								
YINb		Tyrannus	vociferans	2.60	12.22	4.81					5	51	0.72	2.49	0.79															
SineT	55	Tyrannus	verticalis	2.28	14.36	6.36	5.01	1.17	3.82	3.84	1	50	0.62	6.48	1.73	0.13	0.17	0.10	0.19	-0.45	0.86	**1.97**	4.98	-2.12	3.22	**7.65**	0	1	1	0	5
SineT		Tyrannus	vociferans	2.60	12.22	4.81	3.69	1.52	2.75	2.17	5	51	0.72	2.49	0.79	0.26	0.17	0.33	0.22												
Tob	57	Vermivora	chrysoptera	1.66	4.53	2.73	9.50	5.25	6.29	4.24	8	?	0.16	0.54	0.20	0.29	0.26	0.47	0.45	**2.18**	1.00	0.23	0.10	**3.98**	-0.32	-1.94	0	2	0	0	4
Tob		Vermivora	pinus	1.30	3.00	2.42	9.42	3.74	6.52	5.68	10	?	0.16	1.98	1.79	1.02	0.45	0.89	0.90												
YINb	57	Vermivora	chrysoptera	1.72	4.56	2.65	9.11	5.88	6.93	3.23	8	50	0.16	0.58	0.20	0.20	0.47	0.30	0.55	**2.05**	0.87	0.11	**1.80**	0.95	0.67	0.76	1	1	0	0	3
YINb		Vermivora	pinus	1.35	3.22	2.50	8.10	5.36	6.64	2.75	10	50	0.19	1.99	1.74	0.72	0.61	0.53	0.71												
SineT	57	Vermivora	chrysoptera	1.72	4.56	2.65	8.42	5.76	6.76	2.67	8	50	0.16	0.58	0.20	0.37	0.36	0.88	0.50	**2.05**	0.87	0.11	**1.01**	0.70	-0.21	0.40	1	1	0	0	3
SineT		Vermivora	pinus	1.35	3.22	2.50	7.75	5.36	6.94	2.39	10	50	0.19	1.99	1.74	0.83	0.70	0.85	0.80												
Tob	58	Vireo	olivaceus	0.33	2.59	8.14	5.39	2.21	3.32	3.18	9	?	0.06	0.60	1.78	0.34	0.16	0.28	0.31	-1.08	**-1.14**	-0.09	-0.34	**-0.62**	-0.34	-0.17	2	0	0	0	2
Tob		Vireo	philadelphicus	0.40	3.31	8.29	5.56	2.31	3.45	3.25	10	?	0.08	0.66	1.64	0.61	0.16	0.46	0.50												
YINb	58	Vireo	olivaceus	0.37	2.08	6.06	5.10	2.37	3.41	2.73	9	100	0.11	1.12	4.66	0.87	0.17	0.34	0.89	-0.50	**-0.63**	-0.09	-0.41	**-0.84**	-0.75	-0.21	2	0	0	0	2
YINb		Vireo	philadelphicus	0.43	2.68	6.35	5.48	2.56	3.71	2.93	10	100	0.14	0.77	1.37	0.99	0.26	0.44	0.99												
SineT	58	Vireo	olivaceus	0.37	2.08	6.06	5.19	2.26	3.57	2.93	9	100	0.11	1.12	4.66	0.88	0.22	0.46	0.87	-0.50	**-0.63**	-0.09	-0.36	-0.58	-0.59	-0.18	2	0	0	0	2
SineT		Vireo	philadelphicus	0.43	2.68	6.35	5.52	2.42	3.87	3.10	10	100	0.14	0.77	1.37	0.97	0.31	0.53	1.02												

End of Table 5.3.

5.3.4 Discussion

For simple whistles, automatic pitch extraction methods such as YIN-bird and Sine Tracking work very effectively and can greatly benefit zoologists in their analysis. For complex syllables and song it is not as straight forward to extract bird features without the knowledge and supervision of expert listeners who can tell the difference between signature high pitch song and harmonics with F_0 attenuated or missing due to environmental filtering for a given recording.

YIN-bird and Sine Tracking are both sufficient for the task here, with Sine Tracking values slightly more consistent with Tobias et al. A deeper comparison between YIN-bird and Sine Tracking performance on bird song pitch extraction would be an interesting study for future work. The results of experiments here suggest that the difference measure used by Tobias et al. is repeatable using some automatic means. If this system could be fully automated by engineers, it would remove subjectivity when making vocal comparisons while also saving zoologists time by removing the need to visually inspect every spectrogram.

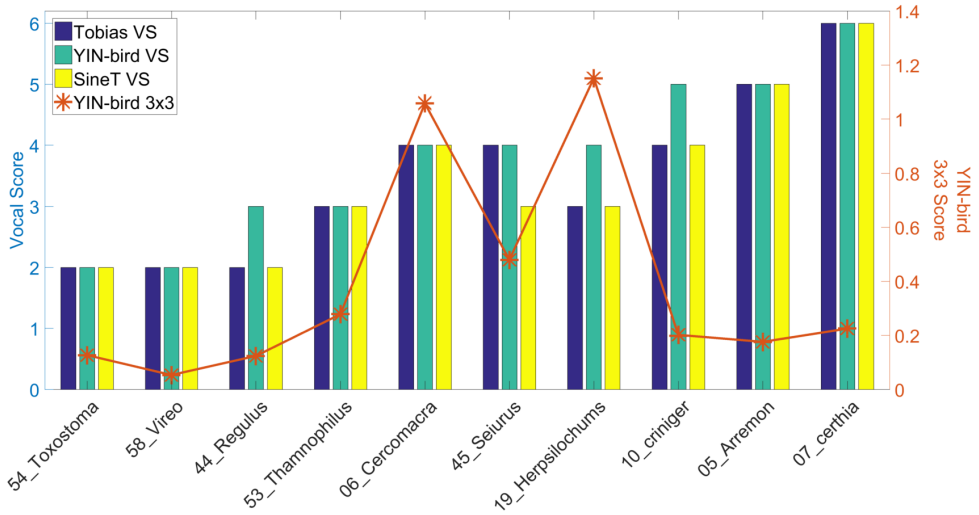
5.4 Comparison: Bird difference measure vs. Tobias measure

A comparison of Tobias et al.'s difference measure and difference measures described in chapter 3 will now be discussed. The reader is reminded that Tobias et al.'s difference measure is influenced by zoological practices, while the latter method utilizes a difference measure inspired by speech processing engineers.

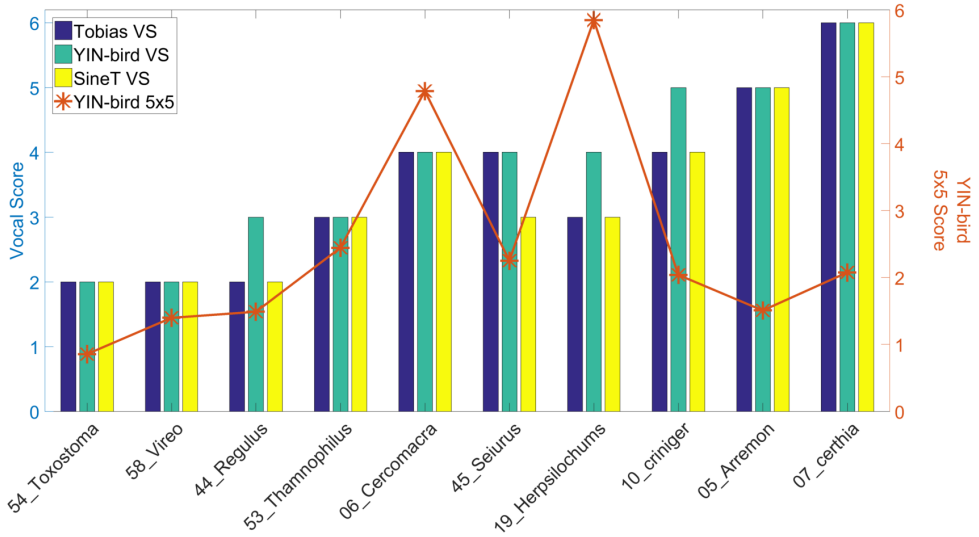
Bird difference scores for this data were done in the same manner as described in chapter 2 and chapter 3. 3x3 and 5x5 codebooks were generated using thresholds set according to conditions: Threshold 1 was chosen so that approximately 70% of ΔF_0 was within $\pm$ Threshold 1 and Threshold 2 was chosen so that approximately 90% of ΔF_0 was within $\pm$ Threshold 2 for each set of populations. Each population pair had its' own specific 3x3 codebook and 5x5 codebook.

Table 5.3 shows vocal scores calculated via manually inspection of spectrograms by Tobias et al. (blue bars in Figure 5.5), scores calculated using segmentation files for temporal features with YIN-bird used for spectral features (green bars in Figure 5.5), and scores calculated using segmentation files for temporal features with Sine Tracking used for spectral features (yellow bars in Figure 5.5). These vocal scores are visually represented by bars in Figure 5.5. Scatter plots with markers superimposed on the bars represent difference scores using 3x3 (Figure 5.5a) and 5x5 (Figure 5.5b) codebooks. Data included in [160, 161], which was annotated by Regan Drennan, but not suitable for analysis using the measures from chapter 3 are listed in Table 5.4. Reasons for inclusion and exclusion are included in the Table also.

What can be learned from Figure 5.5 is that Tobias et al.'s measure and the bird difference measure from chapter 3 do not always agree on the level of difference.



(a) 3x3 codebook comparison.



(b) 5x5 codebook comparison.

Figure 5.5: Bar chart shows vocal scores (VS) from [160,161] compared with VS using YIN-bird and segmentation files. The superimposed scatter plots present the difference score for a given pair using the difference measure from chapter 3.

Table 5.4: Bird pairs excluded from one or both metrics. VS stands for Vocal Score.

Pair No.	Genus	Species	Tobias et al. VS Metric	Ch. 3 Metric
12	Empidonax	hammondi	excluded	excluded
12	Empidonax	oberholseri	Mean freq accurate	Complex vocals
41	Picoides	pubescens	excluded	excluded
41	Picoides	villosus	pitch inaccurate	Nasal vocals
55	Tyrannus	vericalis	excluded	excluded
55	Tyrannus	vociferans	pitch inaccurate	Nasal vocals
57	Vermivora	chrysoptera	included	excluded
57	Vermivora	pinus	Mean freq accurate	Rapid trills vocals

5.4.1 Why difference measures don't always agree

The difference measures appear to disagree for a number of bird pair comparisons. There is some agreement, especially for the 5x5 codebook measure, where the Toxostoma pair has a score of similar and the Certhia pair has a score of different using both metrics. The big disagreements occur for Cercomacra and Herpsilochums pairs that are considered very different using the bird difference measure while using Tobias et al.'s system they are somewhere in between similar and different. To understand why there are inconsistencies, one must consider how each metric is determined.

Tobias et al.'s metric took into account length of song, number of notes and pace (notes/duration) whereas the bird difference metric does not. Tobias et al.'s metric measured minimum frequency, maximum frequency, peak frequency and bandwidth. The bird difference metric of chapter 3 used bi-gram pitch patterns to compare two populations. This involved calculating the probability of going from one code to the next.

Figure 5.6a presents a comparison between song of Cercomacra nigrescens and Cercomacra serva. The spectrogram on the left shows a single song of a Cercomacra nigrescens and the spectrogram on the right shows a single song from a Cercomacra serva. Pitch, extracted using YIN-bird, is superimposed in orange. The mean minimum, maximum and peak frequency values (features used by Tobias et al. but extracted using YIN-bird) for each population is outlined in blue. The features in blue for both populations are similar which leads to a vocal score of 4. In contrast the pitch contours are quite different. The difference measure from chapter 3 exploits the probability of bi-gram (consecutive) pitch codes being observed. Figure 5.6b shows the probabilities of going from code c_j to code c_i with a 3x3 codebook. The matrix on the left corresponds to Cercomacra nigrescens probabilities and the matrix on the right corresponds to Cercomacra serva. Given the overlapping nature of ΔF_{01} and ΔF_{02} some code transitions are not possible. The probability of going from code 9 ($c_j = 9$) to code 4 ($c_i = 4$) is 0.53 for

nigrescens (left matrix, bottom row column 4). The darker the box the higher the probability of that code transition. The reader is reminded that the 3x3 codebook was plotted in Figure 3.2a of chapter 2. The probabilities across a row sum to 1. The highest probability value, for all the data, is constant pitch followed by constant pitch (code 5 followed by 5).

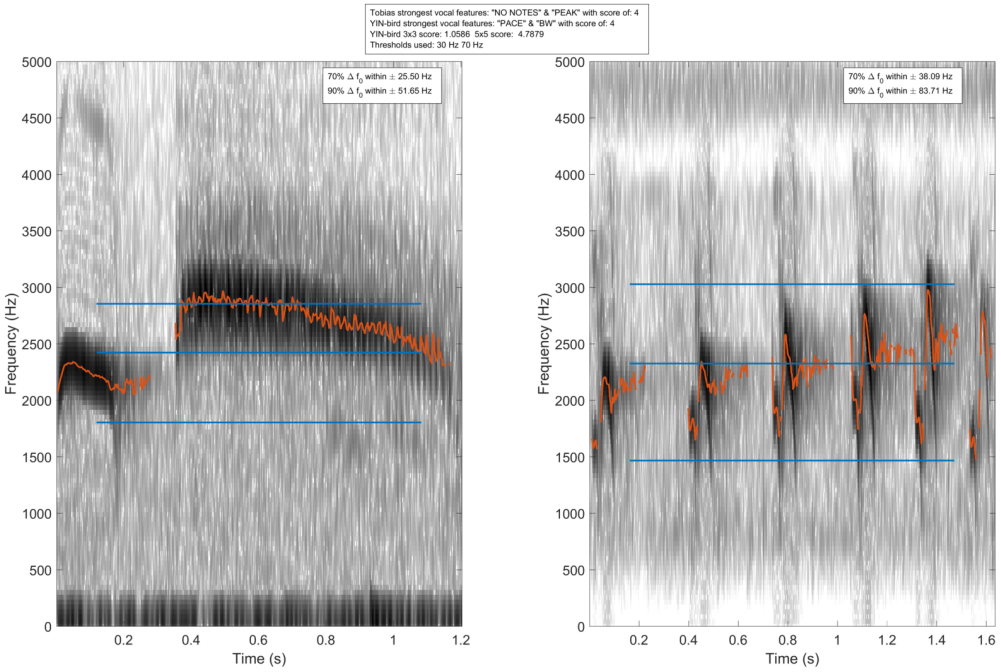
For Cercomacra nigrescens, $P(c_j = 3, c_i = 6) > P(c_j = 3, c_i = 3)$ and $P(c_j = 3, c_i = 6) > P(c_j = 3, c_i = 9)$ which means the pitch contour is more likely to flatten out after rising than continuing to rise or to fall. In contrast for Cercomacra serva, $P(c_j = 3, c_i = 3) > P(c_j = 3, c_i = 6)$ and $P(c_j = 3, c_i = 3) > P(c_j = 3, c_i = 9)$ which means the pitch contour is more likely to continue to rise after rising than flatten out or to fall. This is can be seen in Figure 5.6a where rising pitch contours last longer for serva than for nigrescens song. The same is true for falling pitch, nigrescens song tends to flatten out after falling, while for serva song falling pitch contours last longer (fall for more consecutive frames). So although the bandwidth of their song is similar, the pitch contour structure is quite different. This is why the Cercomacra populations are very different according to bird difference metric and in between similar and different for Tobias et al. score.

Analysis for the Herpsilochmus stictocephalus and Herpsilochmus sticturus displayed in Figure 5.7. This time for Herpsilochmus stictocephalus $P(7, 4) > P(7, 1)$ and $P(7, 7)$ and for Herpsilochmus sticturus $P(7, 7) > P(7, 4)$ and $P(7, 1)$ (see Figure 5.7b). This means that the pitch contour of stictocephalus song is more likely to rise or flatten after falling, whereas the sticturus song is more likely to keep falling in pitch. The spectrogram on the right hand side of Figure 5.7a shows more falling pitch contours than the spectrogram on the left. Again the values of spectral features given by Tobias et al.'s measure are similar, which leads to a lower difference score compared to the score given by bird difference from chapter 3 for the Herpsilochmus pair.

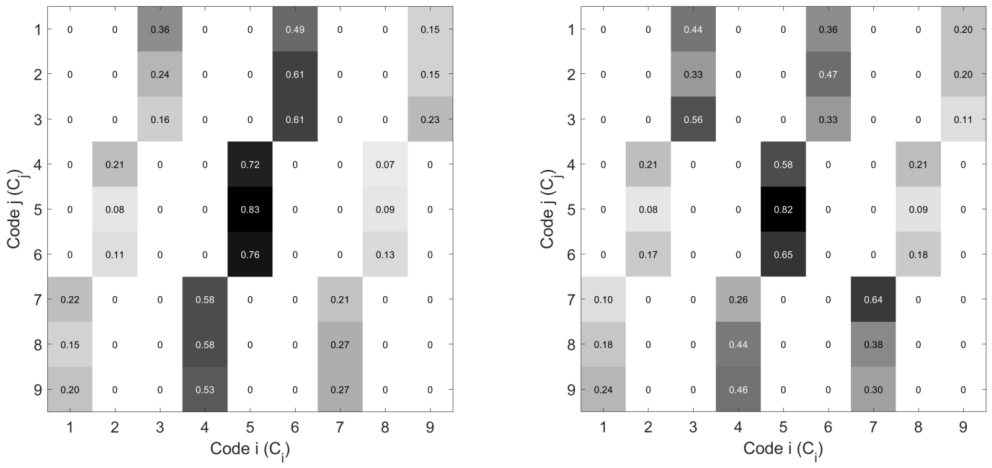
In contrast analysis of the song of Vireo olivaceus and Vireo philadelphicus present low difference scores using both Tobias et al.'s metric and the bird difference metric (see Figure 5.8). Figure 5.8a shows the spectrograms look similar, with Tobias et al. features of similar value (outlined in blue). The matrices in Figure 5.8b show that both populations have similar probabilities of transition from code c_j to c_i.

5.5 Conclusion

This chapter presented automatic means of extracting features from bird song recordings for a standardized criteria of bird species delimitation. Pairs of birds were compared to help quantify what level of vocal difference is required to classify two groups as different species. The reclassification of bird species is a multifaceted task which requires investigation of morphological, visual, acoustic and behavioral characteristics. Work here concentrates on acoustic analysis to aid ornithologists with their work. Current practice involves manual inspection of spectrograms which is tedious and subjective. YIN-bird was utilized to extract spectral features from recordings of bird song. Features automatically extracted from song (together with temporal features

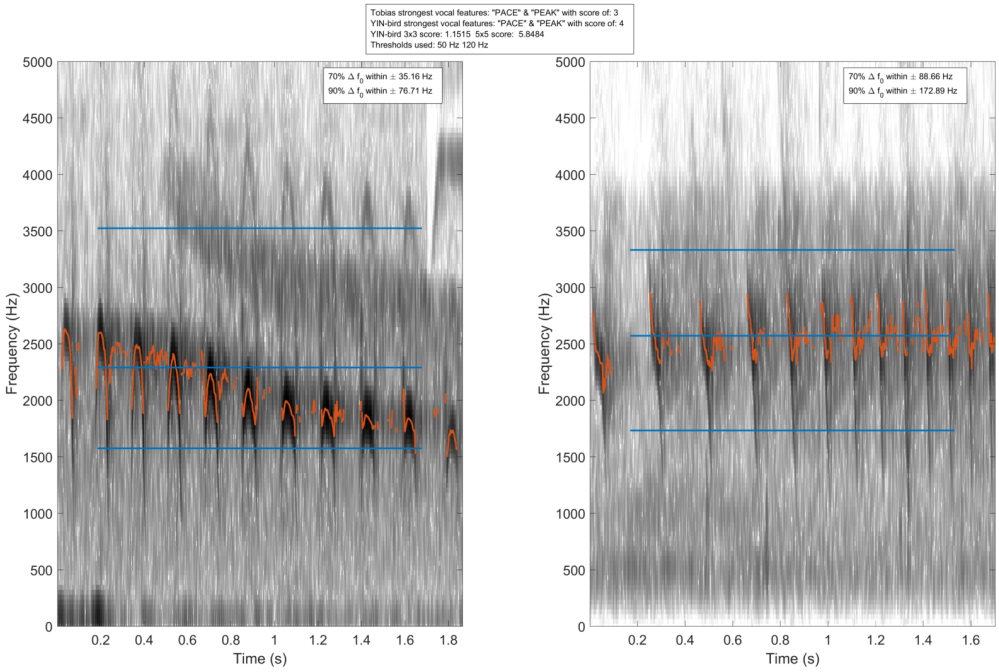


(a) Feature comparison for the song of Cercomacra nigrescens and Cercomacra serva. The spectrogram on the left shows a single song of a Cercomacra nigrescens and the spectrogram on the right shows a single song from a Cercomacra serva. Pitch, extracted using YIN-bird, is superimposed in orange. The mean minimum, maximum and peak frequency values (features used by Tobias et al.) for each population is outlined in blue.

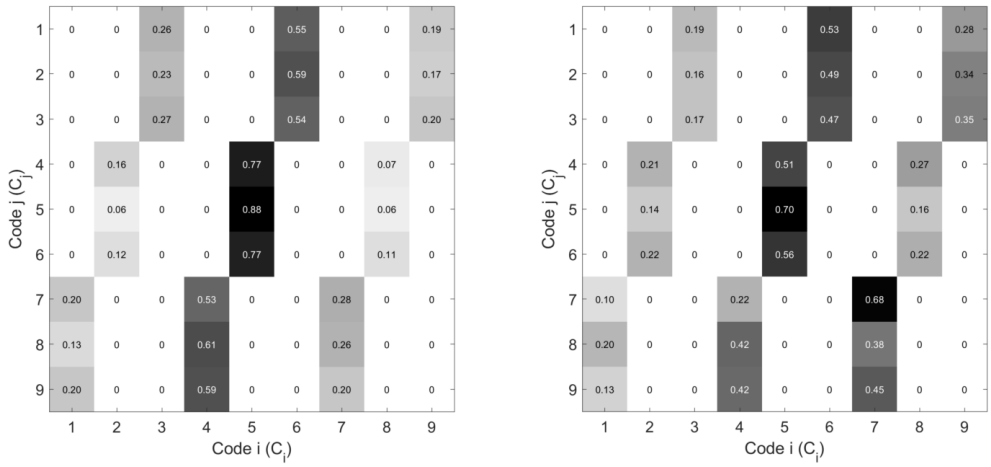


(b) Probabilities of going from code c_j to code c_i. The probability of going from code 9 ($c_j = 9$) to code 4 ($c_i = 4$) is 0.53 for nigrescens (left matrix, bottom row column 4). The darker the box the higher the probability of that code transition. The matrix on the left corresponds to Cercomacra nigrescens probabilities and the matrix on the right corresponds to Cercomacra serva probabilities.

Figure 5.6: Cercomacra nigrescens and Cercomacra serva difference analysis.

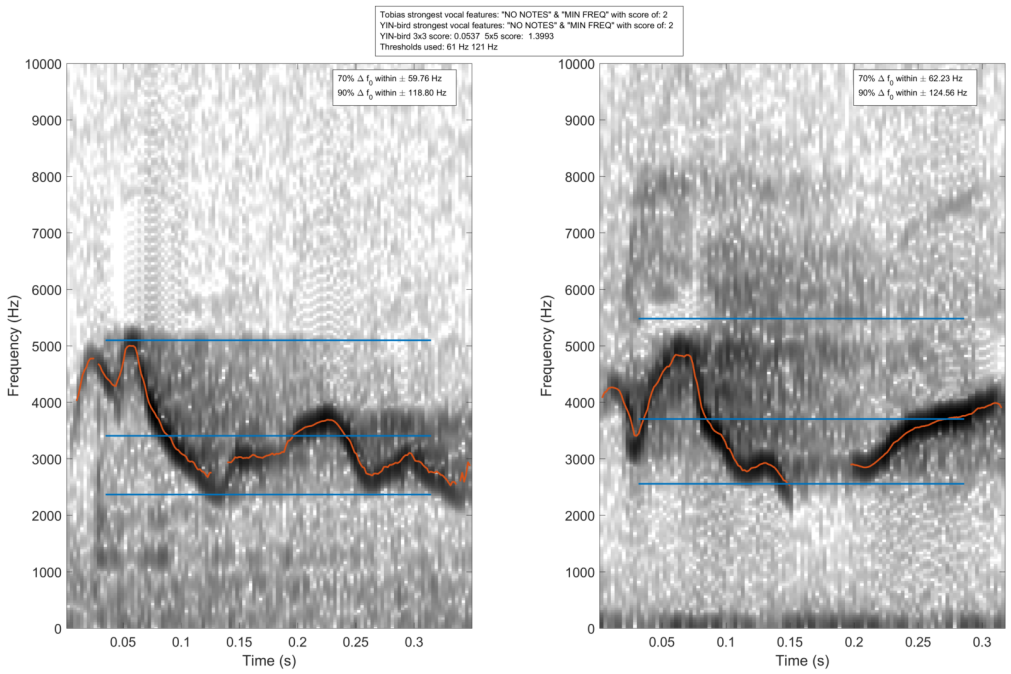


(a) Feature comparison for the song of Herpsilochmus stictocephalus and Herpsilochmus sticturus. The spectrogram on the left shows a single song of a Herpsilochmus stictocephalus and the spectrogram on the right shows a single song from a Herpsilochmus sticturus. Pitch, extracted using YIN-bird, is superimposed in orange. The mean minimum, maximum and peak frequency values (features used by Tobias et al.) for each population is outlined in blue.

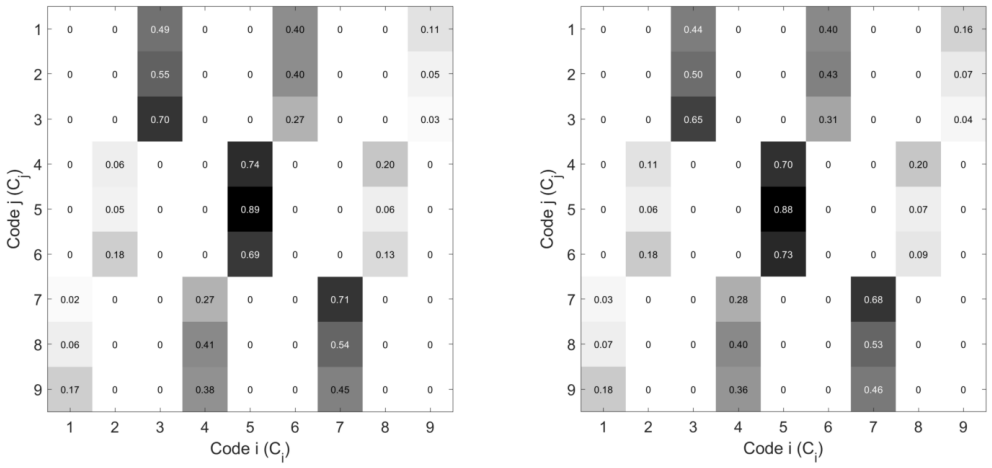


(b) Probabilities of going from code c_j to code c_i. The probability of going from code 6 ($c_j = 6$) to code 5 ($c_i = 5$) is 0.77 for stictocephalus (left matrix, row 6 down, column 5 across). The darker the box the higher the probability of that code transition. The matrix on the left corresponds to Herpsilochmus stictocephalus probabilities and the matrix on the right corresponds to Herpsilochmus sticturus probabilities.

Figure 5.7: Herpsilochmus stictocephalus and Herpsilochmus sticturus difference analysis.



(a) Feature comparison for the song of Vireo olivaceus and Vireo philadelphicus. The spectrogram on the left shows a single song of a Vireo olivaceus and the spectrogram on the right shows a single song from a Vireo philadelphicus. Pitch, extracted using YIN-bird, is superimposed in orange. The mean minimum, maximum and peak frequency values (features used by Tobias et al.) for each population is outlined in blue.



(b) Probabilities of going from code c_j to code c_i. The darker the box the higher the probability of that code transition. The matrix on the left corresponds to Vireo olivaceus probabilities and the matrix on the right corresponds to Vireo philadelphicus probabilities.

Figure 5.8: Vireo olivaceus and Vireo philadelphicus difference analysis.

extracted using manual segmentation) gave vocal scores of similar proportion compared to the scores presented by Tobias et al. in [160, 161]. A comparison between Tobias et al.'s vocal score and the bird difference metric presented in chapter 3 was also presented. Difference scores from both systems for some pairs agreed while other pairs did not. The inconsistencies were explained due to the nature by which each system measures difference. Tobias et al.'s method looks at temporal features and frequency features at a high level while the bird difference metric studies finer pitch contour structures. Songs with similar duration and bandwidth will give a low difference score using Tobias et al.'s method, whereas the bird difference measure will give a high difference score, if the pitch contour structures of these songs are very different. The bird difference metric however does not capture important information like song duration. Both systems have advantages and disadvantages and a combination of both systems may give ornithologists the clearest picture in their evaluations. Future work involves automatic temporal feature extraction which would deal with the difficult task of estimating number of notes per song automatically for bird recordings. Following brief discussions with some zoologists, they believe this system can be deployed in their studies once temporal features can also be extracted automatically.

6

Segmentation of Field Recordings

Despite the wide availability of annotated data for speech, the number of bird recordings with associated annotation information remains relatively small. This is problematic because greater availability of such annotated bird data would improve the accuracy of comparisons between individuals and populations [68].

Many environmental monitoring projects rely on acoustic detection of birds. Despite large datasets, this detection is often manual or semi-automatic, requiring manual tuning/post-processing [148]. This manual pre-processing is time consuming, tedious and subject to the opinion of the expert listeners. This chapter proposes automatic segmentation of bird recordings using bird event detection which identifies song presence with onset and offset time. Such segmentation has important applications in automatic bird species classification, and in the comparison of songs from different populations to aid zoologists in the large-scale study of divergence of species. This is an important step on the long road to making bird song analysis fully automated. In classification, accurate annotation of bird recordings is desirable to ensure that bird species models are built using only vocalizations of the target species and not features extracted from other parts of the recordings. When analyzing bird vocalization difference, accurate segmentation of acoustic events is required, so that subsequent measurements such as bandwidth, peak frequency, minimum and maximum frequency are not affected by non-target signals. Birds' natural habitat is subject to many forms of acoustic competition. Natural sounds from other organisms and human made sounds like industrial noise often corrupt bird recordings.

To apply machine learning to bird vocalizations (e.g. for species classification), data needs to be segmented into phrases, calls or songs. Typically, when working with continuous recordings for

species classification, recordings have been segmented by manual inspection of spectrograms [19]. This can be laborious and prone to subjectivity. Work in [59,116] used data which was segmented into one call per file manually by an expert listener. The data in [71], used for phrase templates, used annotation files containing phrase information labeled by expert listeners. The test data for the HMMs also had to be pre-segmented in that work. The Mexican antbirds data in [165] were divided into one song per file. All of these approaches required some sort of manual pre-processing of data.

Some automatic techniques have been explored for segmentation, e.g. using an energy based threshold to segment the data in time (and sometimes in frequency). An energy based threshold was used in [50, 18, 146]. This approach can be problematic when the amplitude of target and non-target individuals are similar within the same recording. Work in this chapter presents approaches which overcome this limitation. Stowell et al. in [148] discussed a number of different segmentation styles for bird vocalizations. Work here will concentrate on the voice activity detection (VAD) style of segmentation where target/non-target events are identified along with start and end times.

A speaker diarization approach is taken and compared to a baseline segmentation from HMMs in Section 6.2. This is the first presentation that uses a diarization approach to this task, to the best of the authors knowledge. Results are given in Section 6.3. A background to segmentation of bird recordings is detailed in Section 6.1. A discussion of conditions which influence false alarms is given in Section 6.4.

Work in chapter 6 assumes one target bird species per file corrupted by non-target sounds. The start and end times of the target bird's song are desired in what is referred to as automatic segmentation of field recordings. Noisy recordings, for the purposes of this Chapter, refers to recordings of target birds corrupted by non-target vocalizations of other species, recordist speech, traffic, environmental noise or a mixture of all of the above.

6.1 Segmentation of Bird Song

Where annotated data does not exist, signal processing techniques could be used to identify events within a recording (e.g. identify when is a door opening, a person speaking, music playing within a recording). Unsupervised categorization of isolated whale and dolphin vocalizations using DTW and neural networks was discussed in [44]. Machine learning or clustering could also be used to segment longer recordings into target/non-target segments. There has been a lot of interest in isolated bird classification but advances in automatic segmentation or Voice Activity Detection (VAD) for birds is not as well documented. To date signal processing techniques applied to bird vocalizations have concentrated on relatively clean data [51, 55, 57, 59, 145, 165], however this has restricted the analysis of bird populations for which only noisy recordings exist. The ability to accurately and efficiently extract vocalizations from noisy data would widen the range of bird populations which could be investigated.

Automatic bird song segmentation can draw on techniques used in a number of fields. Automatic speaker recognition, source separation, Voice Activity Detection (VAD), event detection and classification, and speaker diarization all incorporate procedures which can help improve bird song segmentation. VAD for speech [53,127,176], is analogous to the task of bird detection in recordings. Time is partitioned into positive and negative regions using energy threshold or model based approaches. Energy based segmentation can be performed by a bioacoustics software tool called Raven [26] and has been used in various surveys [148]. Although energy based time domain segmentation has been successfully applied to low-noise, single-bird recordings, audio from field recordists are generally too noisy for such methods, so a more robust segmentation method is required [109].

A more complex method of VAD uses HMMs. HMMs have been widely used for sound sequence analysis in speech and have particular appeal of temporal flexibility that goes beyond temporal matching [148]. Duan et al. in [48] used Song Scope software [1] which utilizes HMMs for bird detection in a comparison against threshold methods used by Raven. Song Scope is aimed at detecting call structures, which is a different approach to Raven. The Song Scope classification algorithms are based on HMMs using spectral feature vectors similar to MFCCs as these methods have been proven to work effectively in robust speech recognition applications [48, 2]. However, if syllables are polluted by non-target species or background noise, the model is very sensitive, thereby affecting the recognition accuracy. Using Song Scope effectively, requires some background knowledge of signal processing to understand and setup the parameters. Graciarena et al. [57] used a simple VAD system, with acoustic models trained with bird vocalization data to segment their bird recordings. Wei Chu [31] also used HMMs to detect Robin syllable boundaries.

6.1.1 Speaker Diarization

Speaker diarization recovers a structured annotation of time segments which identifies speaker "turns" [164]. Tranter and Reynolds gave an overview of automatic speaker diarization systems in [164]. Audio diarization is defined as the task of marking and categorizing the audio sources within a spoken document. The types and details of the audio sources are application specific. At the simplest, diarization is speech versus non-speech, where non-speech is a general class consisting of music, silence, noise, etc., that need not be broken out by type. A more complicated diarization would further mark where speaker changes occur in the detected speech and associate segments of speech (a segment is a section of speech bounded by non-speech or speaker change points) coming from the same speaker. This is usually referred to as speaker diarization (a.k.a. 'who spoke when') or speaker segmentation and clustering, and is the focus of most current research efforts in audio diarization [164].

There are three primary domains which have been used for speaker diarization research and development: broadcast news audio, recorded meetings and telephone conversations [164].

Most diarization systems have the following component blocks: speech detection, gender and/or bandwidth segmentation, speaker segmentation, speaker clustering and final re-segmentation or boundary refinement [164].

Meignier and Merlin, from the LIUM group in France, developed an open-source speaker diarization toolbox in Java, mostly dedicated to broadcast news and made it available at the 2010 Sphinx Users and Developers Workshop [104]. The toolkit, entitled 'LIUM_SpkDiarization', was reviewed and improvements were discussed in [133] at INTERSPEECH 2013.

Speaker diarization, the "who spoke when" task, consists of annotating recordings with labels representing the speakers. This task is performed without any prior information: neither the number of speakers, nor their identities, nor samples of their voices are available [133]. The technique used by LIUM of identifying acoustic changes and clustering events between the change boundaries is exploited and applied to bird song recordings in this Chapter.

Maina [168] was the first to apply speaker diarization techniques to bird song analysis. In [168], speaker diarization was used to identify change points and cluster segments into classes. One class was assumed to be non-bird sounds, while the rest of the clusters were taken to indicate the number of species present in the recording. Thus the system explored the relationship between clusters and species counts in recording of 1-6 birds. The correct number of species was obtained for 76 of 179 recordings. Work here utilizes a similar speaker diarization setup but instead now applied to bird song segmentation. The ability to identify change points throughout a recording is central to the potential strength of such an approach. Experiments evaluate segmentation accuracy of target song (with time boundaries) from field recordings with a single target bird corrupted by other audio events. Trained models can refine boundaries and remove non-bird events detected as song. The system is compared to baseline segmentation using HMMs.

6.2 Experimental Setup

Segmentation was treated as a two class recognition problem. Each recording was assumed to have a dominant target bird vocalizing, with varying levels of background or non-target interference present in the same recording. The segmentation used here was at the song level of annotation which means any information between the start and end of a song bout was labeled as song. Song can contain many syllables which are separated by silence or run into each other. Information contained outside these bouts of song were labeled as background or 'Bg'. Bg contained signals from other sources such as insects, other bird species, humans and environmental noise. Bg signals overlap with target song from time to time. In these cases if song is present, it is labeled as song even when non-target corruption is high. Zoologists require segmentation of relevant parts of recordings, i.e. extract target song and reject calls and non-target audio from a recording. This is a complex problem as calls from one species may sound more like song of another species. Automatically extracting target vocalizations is the first step in a bigger project. All target audio in this data set was song. If the data contained target calls

it should be labeled as target song for these experiments.

Figure 6.1a shows a clean recording example. The top panel shows the time domain signal, middle panel shows the spectrogram and the bottom panel represents song level annotation. The white regions either side of the area labeled 'Song', are considered 'Bg'. The full file is shown in Figure 6.1b; note that at this zoom setting, the word 'Song' is hidden and/or displayed as 'S' between blue boundary lines.

In Figure 6.1c the bird song is not as clear in the time domain. The spectrogram shows the song corrupted by insect sounds. Figure 6.1d shows speech at the start of the recording, with a similar amplitude to the bird vocalizations (speech is highlighted in pink in the time domain part of Figure 6.1d). In cases like this, amplitude threshold segmentation will fail because the energy of the speech is closer to the energy of bird vocalizations than it is to the noise floor energy level. Figure 6.1e shows a difficult case where two different species are vocalizing at similar amplitudes. As the characteristics are quite similar, segmentation is problematic unless segmentation was done using trained models for all species, like species classification. Currently it would not be practical to train a model for every species that may be corrupting target birds for a given dataset.

Figure 6.1f shows an example corrupted by noise, insects and a non-target bird. Figure 6.1g show an example of a recording with multiple birds vocalizing with overlap. There is also some wideband noise.

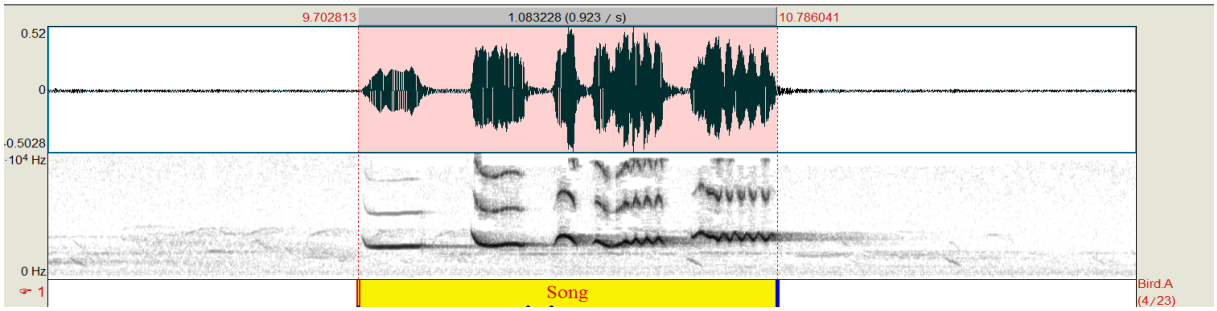
These examples are shown to give the reader an example of the different quality types and problems which occur within the dataset. The recording conditions within the dataset are highly variable.

6.2.1 Data for Segmentation

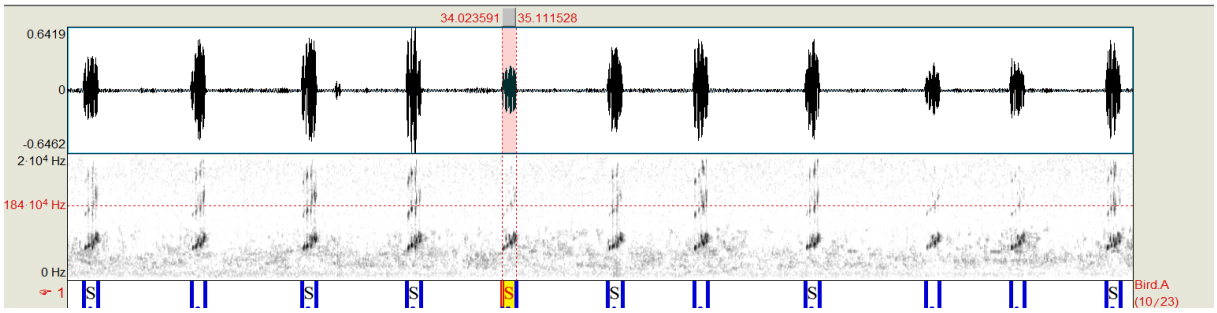
The data used for experiments in this chapter comes from the same dataset used in chapter 5. In chapter 5, labeled bouts of song were extracted from long recordings. Due to time constraints, sometimes only a few songs were labeled within a long recording. Typically clear examples and stereotypical songs were labeled and poorer examples ignored. Experiments in chapter 5 involved analyzing song. These bouts of song were extracted from longer recordings using manual annotation. Experiments in this chapter establish automatic means of annotating long recordings and thus using the automatic annotation to automatically extract bouts of song. These experiments required annotations of whole recordings for training and evaluating test results from ground truth. Therefore only recordings with annotations for the whole file were used here.

The annotation process was described in detail in Section 5.3.2 of chapter 5. A list of data used here is given in Table 6.1. The sources of data are the same as given in Table 5.2. The average length of whole recordings, per species, are given in the last column.

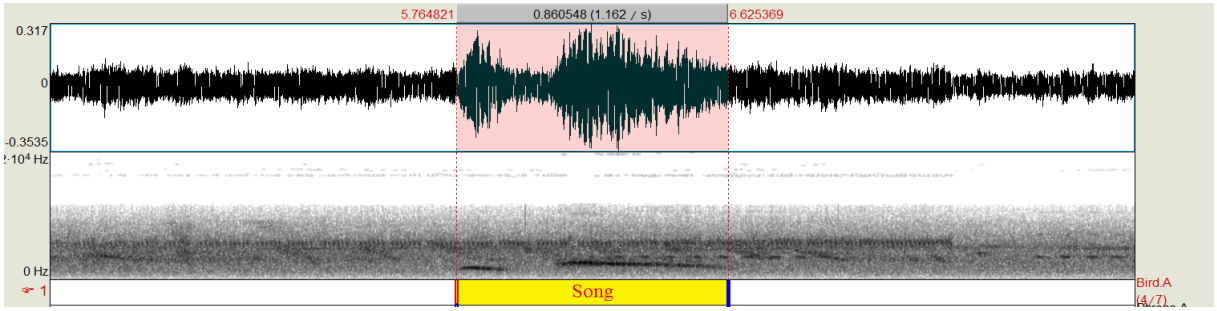
As there were only 44 files used in these experiments (see Table 6.1), cross-validation was



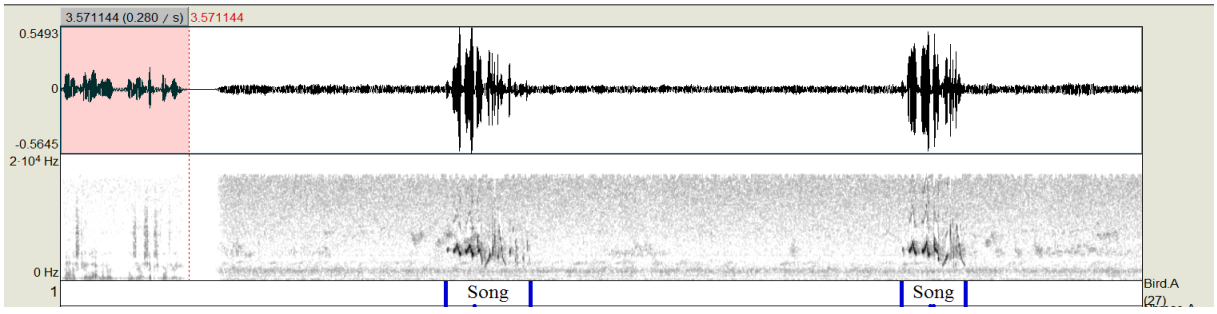
(a) Certhia brachydactyla *Praat* screenshot, showing clean recording with song annotation.



(b) Certhia brachydactyla *Praat* screenshot, showing full recording with 11 songs annotated.

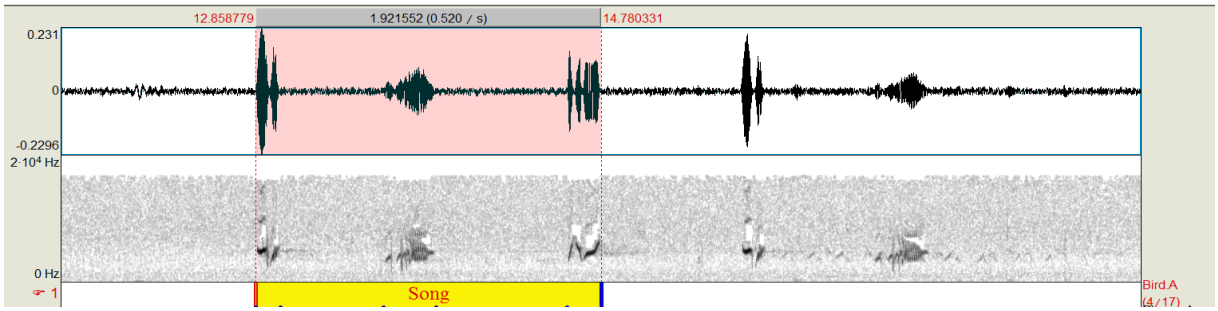


(c) Cercomacra nigrescens *Praat* screenshot, showing insect sounds corrupting the bird recording.

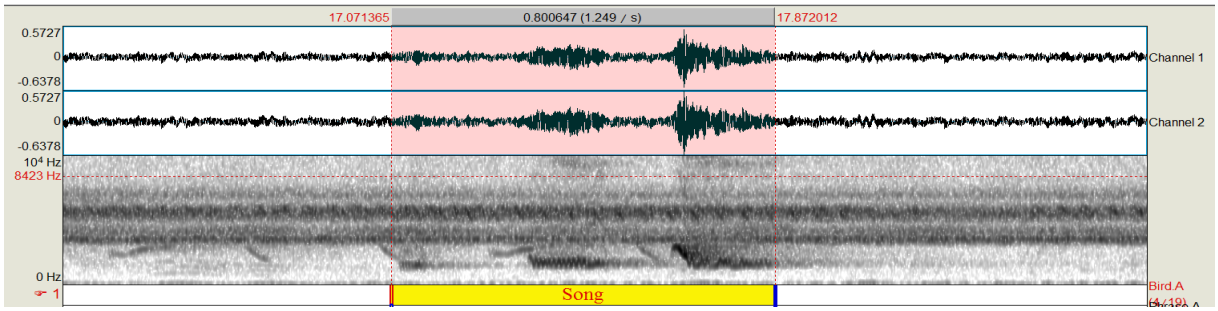


(d) Seiurus motacilla *Praat* screenshot, with speech at the start of the recording (highlighted in pink).

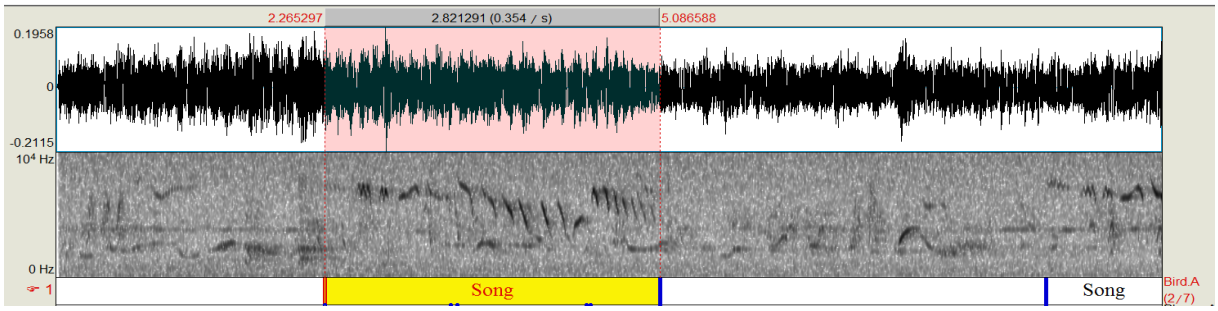
Figure 6.1: Continued on next page ...



(e) Empidonax oberholseri *Praat* screenshot. Two different species birds are singing close together.



(f) Criniger ndussumensis *Praat* screenshot, with insects, environment noise and another non-target bird.



(g) Certhia familiaris *Praat* screenshot, with wideband noise and other birds singing.

Figure 6.1: Screenshots of *Praat* showing time (top) and frequency (middle) domain information along with labeling of song boundaries (bottom).

used where 43 files were used for training and one for testing for each split. Results were concatenated into a single result vector so mean evaluation scores take into account the length of each test file.

6.2.2 Evaluation

To gauge the segmentation accuracy, an evaluation metric used for VAD was implemented. The performance of the segmentation system was assessed by comparing percentage of correct classifications (song and background) with manual annotations for 44 test files. The performance metrics used in [11] were percentage of correct speech identification (*PCS*) , percentage of correct

Table 6.1: Data for Segmentation.

No.	Common Name	Genus	Species	No. of Files	Mean Dur (s)
1	Chestnut-capped Brush Finch	Arremon	brunneinuchus	3	23.63
2	White-browed Brush Finch	Arremon	torquatus	3	36.57
3	Blackish Antbird	Cercomacra	nigrescens	4	26.47
4	Black Antbird	Cercomacra	serva	4	36.37
5	Short-toed Treecreeper	Certhia	brachydactyla	3	67.25
6	Eurasian Treecreeper	Certhia	familiaris	3	44.11
7	Red-tailed Greenbul	Criniger	calurus	2	83.50
8	White-bearded Greenbul	Criniger	ndussumensis	2	62.75
9	American Dusky Flycatcher	Empidonax	oberholseri	1	56.34
10	Todd's Antwren	Herpsilochmus	stictocephalus	3	40.12
11	Spot-tailed Antwren	Herpsilochmus	sticturus	1	36.04
12	Common Firecrest	Regulus	ignicapillus	3	114.88
13	Goldcrest	Regulus	regulus	3	63.35
14	Louisiana Waterthrush	Seriurus	motacilla	4	126.52
15	Northern Waterthrush	Seiurus	noveboracensis	5	60.01
			Total	**44**	**2,612**

non-speech identification (*PCN*) and percentage of false identification (*PF*). For experiments here, the word speech refers to bird song and non-speech to background sounds. The Matlab function 'hrate' by Jamal Saeedi [135] was used to perform these measures.

All the recordings were made at $\boldsymbol{F_s}$ of 44.1 kHz. The ground truth annotation times were converted into a binary vector (vector length (N) = 44100 * length of test file in seconds). Song was specified as a one and background as a zero. The output label times given by the segmentation system in use, were also converted into a binary vector. The ground truth and output vectors were compared using Eq. 6.1, 6.2 & 6.3 at the sample level.

$$PCS = 100\ [mean(result_vector(GT_vector == 1))] \tag{6.1}$$

$$PCN = 100\ [1 - mean(result_vector(GT_vector == 0))] \tag{6.2}$$

$$PF = 100 \sum_{n=1}^{N} \frac{|\ GT_vector - result_vector\ |}{length(GT_vector)} \tag{6.3}$$

where result_vector is a binary sequence of system output predictions with one for 'Song' and zero for 'Bg'; and GT_vector is a binary sequence specifying ground truth labels with one for

Table 6.2: MFCC parameters for HTK experiments.

Parameter	Value
SOURCEKIND (I/p file)	WAVEFORM
SOURCEFORMAT	WAV
TARGETKIND (O/p file)	MFCC_E
WINDOWSIZE	25 ms
TARGETRATE (Frame rate)	10 ms
USEHAMMING (T = True)	T
NUMCHANS (No. mel filters)	25
CEPLIFTER (Cep liftering)	0
NUMCEPS (No. of MFCC's excluding E)	12
ENORMALISE	T
LOFREQ (Start freq)	0 Hz
HIFREQ (Cut-off freq)	22050 Hz

*'Song' and zero for 'Bg'. Both vectors are of length N, with N = $\boldsymbol{F_s}$ * length of audio file in seconds. Equation definitions: percentage of correct speech identification (PCS), percentage of correct non-speech identification (PCN), percentage of false identification (PF) [135, 11].*

6.2.3 Segmentation using Hidden Markov Models (HMMs)

HMMs with between 3 and 8 states were trained using HTK version 3.4 [175]. The number of mixtures per state was gradually incremented to 64. The parameters used are given in Table 6.2. 12 MFCCs with energy were chosen as features for these experiments because of their successful application in bird species recognition, speech and speaker recognition, acoustic event detection, and other technologies. Previous work with MFCCs was reviewed earlier in this chapter and in chapter 2. To perform segmentation using HMMs, various tools from speech recognition toolkit HTK [175] were used. MFCC features were extracted using 'HCopy'.

After feature extraction, HMM models for 'Song' and 'Background' were trained. 'Song' was specified as a left to right model and 'Background' was an ergodic model. For the first stage of training, 'HCompV' was executed. This program calculates the global mean and covariance of a set of training data. It is primarily used to initialize the parameters of an HMM such that all component means and all covariances are set equal to the global data mean and covariance [175]. Next initial estimates for the parameters of a single HMM using a set of observation sequences were computed using 'HInit'. After initialization, 'HRest' performed basic Baum-Welch reestimation of the parameters of a single HMM using a set of observation sequences. This was performed five times to optimize parameters. At this stage, models had a single Gaussian per state. 'HHed' manipulates sets of HMM definitions. Its basic operation is to load

in a set of HMMs, apply a sequence of edit operations and then output the transformed set [175]. This program was executed to edit each model so that there were 'M' mixtures per state. After 'HHed' another round of 'HRest' was executed to optimize model parameters with an updated number of mixtures. The number of mixtures was increased gradually, $m = 1, 2, 4, 8, 16, 32, 64$, and 'HRest' was performed in between each number of mixtures increase. The EM algorithm is sensitive to good initialization and prone to converging to local maxima. The multiple calls of 'HRest' and slowly increasing the number of mixtures reduces the risk of local convergence during the training phase.

During evaluation, long recordings were used as test data. Features were extracted in the same manner as for training. 'HParse' was used to generate a lattice file which dictates that predictions can be either 'Song' or 'Background' no matter what model the previous instance prediction was. 'HParse' is a tool from HTK which builds a word network for HMMs in speech recognition. Next 'HVite' is used to output the system's predictions using Viterbi. 'HVite' is a general-purpose Viterbi class recognizer. It will match an input file against a network of HMMs and output a transcription for each [175]. The results were stored in a .mlf (master label file) and were evaluated by comparing results to ground truth segmentation using the methods described in Section 6.2.2.

6.2.4 Segmentation using speaker diarization tool LIUM

LIUM [104, 133] is an open-source speaker diarization toolbox, designed with to be applied to broadcast news predominantly. The toolkit includes HAC using measures such as Bayesian Information Criterion (BIC) and Cross Likelihood Ratio (CLR). Experiments here evaluate LIUM's use on bird recordings, specifically for the task of segmentation.

LIUM comes with Sphinx4 built in. Sphinx4 was used to extract the features for these experiments. A linear filterbank was used instead of a Mel spaced one, as the frequencies of interest in bird song are higher than speech. Intuitively a linear filterback would be more suitable than Mel spaced, but segmentation accuracies showed little difference between the two front-ends in the development phase. HTK doesn't have a linear filterbank option hence why mel spacing was used. The parameters used are shown in Table 6.3. The same dataset used for HMM experiments was used for LIUM experiments.

The first stage involved extracting features from a test recording to be segmented. Cepstral coefficients for periods of a recording which contains no energy (e.g. when microphone was switched off) should transform into a vector of zeros, but a bug with LIUM hardcoded the features to large numbers which caused models to return NaNs as centers. To debug this, features ≥ 1000 were zeroed. The test files feature vector was passed through 'MSegInit' which performs basic safety checks (checks the feature vectors to ensure that there are no sequences of several identical vectors and checks that the sections of the file on which segmentation is to be done fits) and 'MSeg' which is a similarity based segmentation detection program that

Table 6.3: Feature parameters for LIUM experiments.

Parameter	Value
features	Linear_CC_E
windowSizeInMs	25 ms
windowShiftInMs (Frame rate)	10 ms
hamming (T = True)	T
numberFilters	25
premphasizer	0.97
no. static coefficients	12
energy coefficient	1
total size of a feature vector	13
normalization	None
minimumFrequency	0.33 Hz
maximumFrequency	22000 Hz

finds instantaneous change points corresponding to segment boundaries. Detection performed by 'MSeg' at that point was done using Generalized Likelihood Ratio (GLR) distance. 'MSeg' detects the change points through a similarity, computed using two Gaussians. The Gaussians are estimated over a window sliding along the whole signal. Initial change point, i.e. a segment boundary, is present in the middle of the window when the similarity GLR reaches a local maximum. Firstly a lot of change points are assigned, as it favors over-segmentation. The next pass used 'MClust' to linearly compare neighboring segments with parameters '-l' and '–cThr=2' (Penalty factor in similarity). BIC distance between segments was used to fuse consecutive segments that were found to correspond to the same class. 'MClust' was run a second time with parameters changed to '-h' (hierarchical rather than linear) and '–cThr=3'(the threshold used for stop criterion now). This time, the HAC merged the two closest clusters at each iteration until the best BIC distance was positive. In this step and in the previous one, classes were modeled by a Gaussian distribution with a full covariance matrix. BIC distance, ΔBIC, is defined in Eq. 6.4 and penalty factor, P, is defined by Eq. 6.5.

$$\Delta BIC_{i,j} = \frac{n_i + n_j}{2} \, log|\Sigma| - \frac{n_i}{2} \, log|\Sigma_i| - \frac{n_j}{2} \, log|\Sigma_j| - \lambda P \tag{6.4}$$

$$P = \frac{1}{2}\left(d + \frac{d(2+1)}{2}\right) + log(n_i + n_j) \tag{6.5}$$

where $|\Sigma_i|, |\Sigma_j|, |\Sigma|$ *are the determinants of Gaussians associated to the clusters* i, j *and* $i + j$. λ *is a parameter specified by user.* P *depends on* d*, the dimension of the features, as well as on* n_i *and* n_j *referring to the total length of cluster* i *and cluster* j *respectively. The two closest clusters* i *and* j *are merged at each iteration until* $\Delta BIC_{i,j} > 0$

Table 6.4: HTK HMM experiments' percentage of false identification (*PF*) for varying number of mixtures and states.

	No. of Mixtures:					
No. of States:	**2**	**4**	**8**	**16**	**32**	**64**
3	22.81	21.74	23.69	22.15	24.10	19.80
4	19.39	19.42	19.43	21.81	21.99	24.90
6	22.50	19.38	19.63	**19.20**	25.38	26.56
8	25.42	24.12	25.10	20.68	20.59	21.72

Using the clusters up to this point, GMMs were trained for each class (speaker when dealing with diarization, clusters of target song here). The GMMs were trained using diagonal covariance by 'MTrainInit' and 'MTrain' (referred to as 'Test' models). The initialization method was optimized using '-emInitMethod=split', which split the Gaussian with biggest variance into two new Gaussians at each iteration until the number of Gaussian components was met). These models were used to generate new segments for the recording using 'MDecode', a Viterbi decoder which utilizes GMMs. Segment boundaries were adjusted using 'SAdjSeg' next.

Work to this point required no training. When performing speaker diarization with LIUM, 512 component GMMs for speech, non-speech, silence and gender Universal Background Models (UBMs) are used. For segmentation of bird recordings, the 43 recordings in a given training split were used to train a 'Song' UBM and a 'Background' UBM. The number of mixtures for these models was varied from 28 to 80. These GMMs were used by 'MDecode' which performed Viterbi decoding to create a segment file based on these models, which identified regions of background audio and song. This segment file was then used as a filter mask to make refinements to the HAC segment output by 'SAdjSeg', using 'Sfilter'. Finally 'SSplitSeg' was used to make boundary refinements such as split segments longer than 20 ms (which is not really needed for bird segmentation but useful for speaker diarization). This process generates a segment file which identifies instances of song and their temporal boundaries. The songs are generally clustered into one to five "Speakers" (even if only one bird was present). This is discussed in Section 6.4.2. As the goal of this work was to identify song boundaries within a long recording and not to perform speaker diarization for birds, all cluster labels were transposed to label 'Song'. This segment file which identified song boundaries was then compared to the ground truth using the evaluation metric presented in Section 6.2.2.

6.3 Results

The percentage of false identification (*PF*) scores (for which higher value means higher error) for each setup of HTK HMM experiments are provided in Table 6.4. The optimum *PF* was found with 6 states and 16 mixtures per state for these experiments.

Table 6.5: LIUM experiments' percentage of false identification (*PF*) for varying number of mixtures for train models and test models.

	No. of Train Mixtures:									
No. of Test Mixtures:	**28**	**32**	**36**	**40**	**48**	**56**	**64**	**70**	**75**	**80**
6	19.67	16.00	15.12	15.32	15.32	16.17	15.09	-	-	-
8	19.67	15.99	15.11	15.33	15.33	16.18	15.08	**15.01**	15.11	15.40
12	19.68	16.00	15.12	15.35	15.34	16.19	15.10	-	-	-
16	19.70	16.02	15.14	15.37	15.37	16.21	15.12	15.05	15.15	15.43
20	19.69	16.02	15.14	15.37	15.37	16.20	15.12	-	-	-
24	19.67	16.00	15.12	15.35	15.36	16.20	15.10	15.03	15.13	15.43

Table 6.6: Segmentation percentage of correct speech identification (*PCS*), percentage of correct non-speech identification (*PCN*) and percentage of false identification (*PF*) percentages for optimal setup of experiments.

Experiment	***PCS*** **(%)**	***PCN*** **(%)**	***PF*** **(%)**
HTK	87.06	79.01	19.20
LIUM	86.41	84.58	15.01

The results of LIUM segmentation experiments can be seen in Table 6.5. Here LIUM parameters were varied. The number of mixtures for the 'Song' and 'Background' UBMs were varied from 28 to 80. The number of mixtures per test model varied from 6 to 24. Test models refer to models trained at a stage of LIUM when a GMM is trained for each early cluster within the test recording. These test models are used to refine segments. Therefore if there were 5 early stage clusters within a recording each of the 5 clusters would have 6, 8, 12, 16, 20 or 24 mixtures for a given experiment (i.e. each cluster regardless of data size has the same number of mixtures). The number of mixtures for test models should be lower as each GMM model has less data than the UBMs, and one does not want to overfit the test models. The best LIUM score was observed using 70 mixtures for train models (UBM) and 8 mixtures for test models (clusters). The accuracy for LIUM is shown to be dependent on the number of train mixtures but varying the number of test mixtures has little effect on the final *PF* value. This is because test models typically have 5 clusters for a 60 second recording. There is not enough data per model to utilize the increase in mixture numbers. Best performance was found around 8 mixtures.

Table 6.6 gives the percentage of correct speech identification (*PCS*), percentage of correct non-speech identification (*PCN*) and percentage of false identification (*PF*) percentages for the optimal setup of both the LIUM and HTK HMM systems. For HTK HMMs with 6 states and 16 mixtures, $PCS = 87.06\%$, $PCN = 79.01\%$ and $PF = 19.20\%$. For a speaker diarization

system from LIUM using GMMs with 70 mixtures for training models and 8 mixtures for test models, *PCS* = 86.41%, *PCN* = 84.58% and *PF* = 15.01%. Note *PCS* and *PCN* are percentage accuracies and *PF* is a percentage error.

The improved performance in *PCN* for LIUM over HTK HMMs suggestions less false alarms for non-target song being identified as song. This is the main benefit of LIUM, it has a greater ability to reject background/non-target song than HTK HMMs.

6.4 Discussion

Segmentation using LIUM is attractive because it satisfies the need for a general purpose segmentation approach that generalizes well to a variety of situations, recording conditions and species. Speaker diarization has been shown to successfully segment broadcast news without knowing the number of speakers and without prior samples of their voices [133]. For bird segmentation the goal is to effectively segment song from a recording by identifying change points. Further refinements are then made by a background model of bird song and a model of non-bird song. The species under test does not have to be used for training, it is assumed that song from an unseen target will match the bird song model more than the non-bird model. This is also true for segmentation using HTK HMMs however the boundary detection will not be as accurate as change points are not compared.

Figure 6.2 shows the segmentation of a recording of Seiurus motacilla. The figure from top to bottom is laid out as follows: spectrogram, expert listener annotations, HTK HMM automatic segmentation result and LIUM automatic segmentation result. A blue area surrounded by red border corresponds to '1' or song label. White area in between blue rectangles corresponds to '0' or background label. Upon listening to the recording faint background chorus from other birds can be heard. This is incorrectly labeled target song at instances of the HMM result, as shown at regions where the HMM song/bg shows blue while the ground truth labeled by expert listeners shows white for background. The LIUM segmentation plot is more consistent with the ground truth. The end of the recording contains speech by the recordist and both automatic methods mistake speech for song. This is one of a few examples where speech is mistaken for song. For this example, HMM *PF* was 22.94% and for LIUM *PF* was 3.37%. LIUM rejects non-target bird sound better than HMMs in this example.

Figure 6.3 shows segmentation of a recording of Regulus regulus. HMMs incorrectly segmented the single bout of song as being shorter than it truly was, LIUM did not make the same mistake.

Figure 6.4 shows an extreme example where HTK HMM segmentation fails. The recording of Herpsilochmus stictocephalus contains other background chorus song and insect noise. HTK HMMs output five song segments with negligible pauses between them. LIUM performs better but makes a mistake at labeling a non-target song as target song around 17 seconds, and there was an insertion error of a pause around 25 seconds. For this example, HTK HMM *PF* was

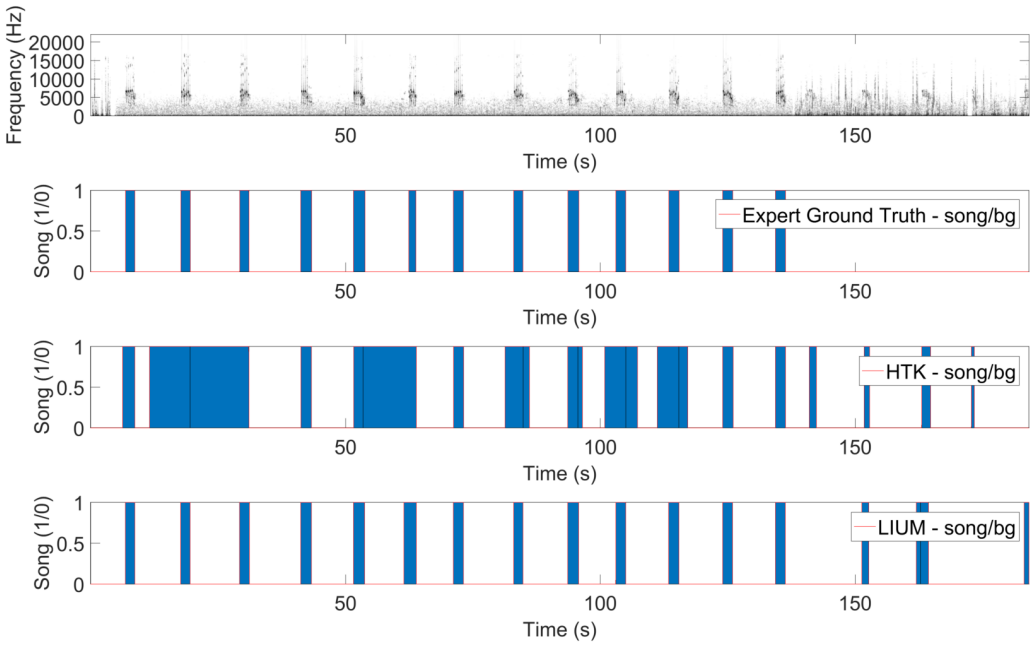


Figure 6.2: Seiurus motacilla recording spectrogram (top subplot), ground truth annotation by expert listener (second subplot), HTK HMM segmentation result (third subplot), LIUM segmentation result (bottom subplot). Blue area corresponds to song label, white corresponds to background or non-target label. For this example, HTK HMM *PF* was 22.94% and for LIUM *PF* was 3.37%.

88.93% and for LIUM *PF* was 10.71%. The non-target song, when listened to, sounded not dissimilar to the target song. A human listener might mistake example as song, LIUM mistakes are akin to mistakes non-expert human listeners would make while HTK HMMs mistakes insect sounds for song, a mistake which even inexperienced listeners would not make.

Figure 6.5 shows an example of Certhia familiaris. This is an example where HTK outperforms LIUM. The recording contains a target bird with non-target birds signing with similar amplitude. LIUM inserted two extra song labels and overestimated the duration of the first song. For this example, HTK *PF* was 6.84% and for LIUM *PF* was 30.21%. When listening to this recording it was quite difficult to tell which bird was which. Perhaps HTK performs well simply because the target bird song is at a higher frequency and these frequencies better matched the song HMM model. LIUM has difficulty identifying boundaries when overlap between target and non-target bird song is observed.

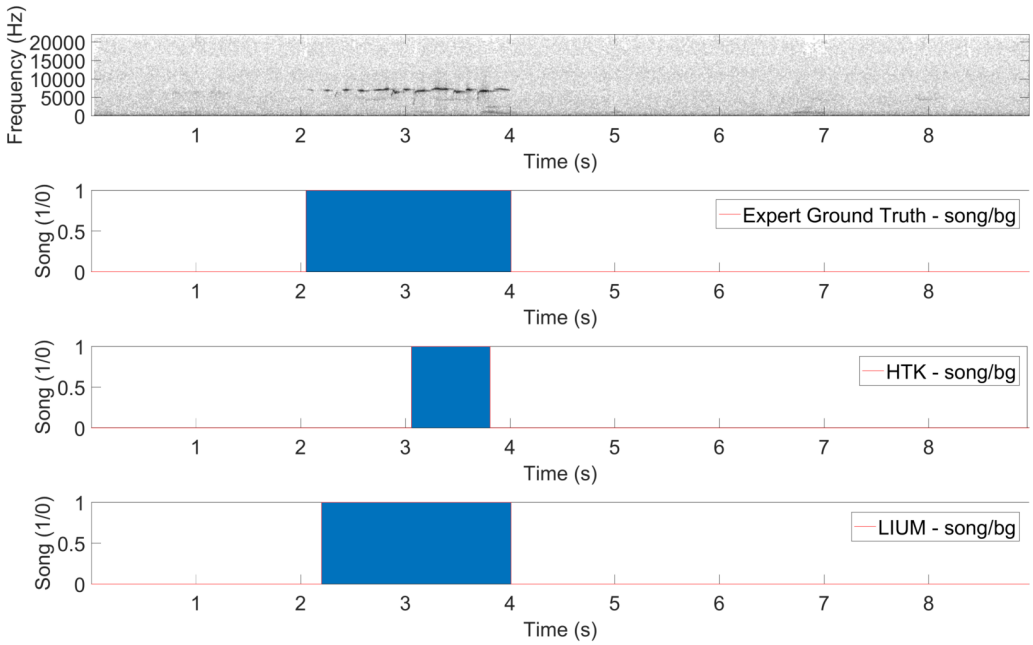


Figure 6.3: Regulus regulus recording spectrogram (top subplot), ground truth annotation by expert listener (second subplot), HTK HMM segmentation result (third subplot), LIUM segmentation result (bottom subplot). Blue area corresponds to song label, white corresponds to background or non-target label.

6.4.1 Qualitative discussion of segmentation

By inspecting results from all 44 splits trends emerge. The HMM system tends to identify background bird song as song even when amplitude value suggests song is distant. LIUM correctly rejects these types of non-target song, as although change boundaries identify these regions as song initially, such clusters are identified as non-target segments during the 'Sfilter' step and removed. Medium noise conditions (qualitatively medium) prove difficult for MFCCs in pattern classification, hence the HMM system often fails, e.g. false alarm song when it should be background. LIUM however still identifies change points in these conditions which allows segmentation with higher accuracy. In more severe Signal-to-Noise Ratio (SNR) conditions, LIUM still outperforms HMMs, as detecting change points is advantageous over identifying noisy frames when comparing them to trained models.

A major factor for false alarms with LIUM is non-target birds sounding like target birds, e.g. cases where non-expert listeners would mistake the non-target song for target song. HMMs are also prone to this type of error. Sometimes in noisy examples, HMMs performed better than LIUM. Non-target birds in presence of noise were not detected as song by the HMM system but LIUM identified them as change points (usually with a different cluster) with false alarm

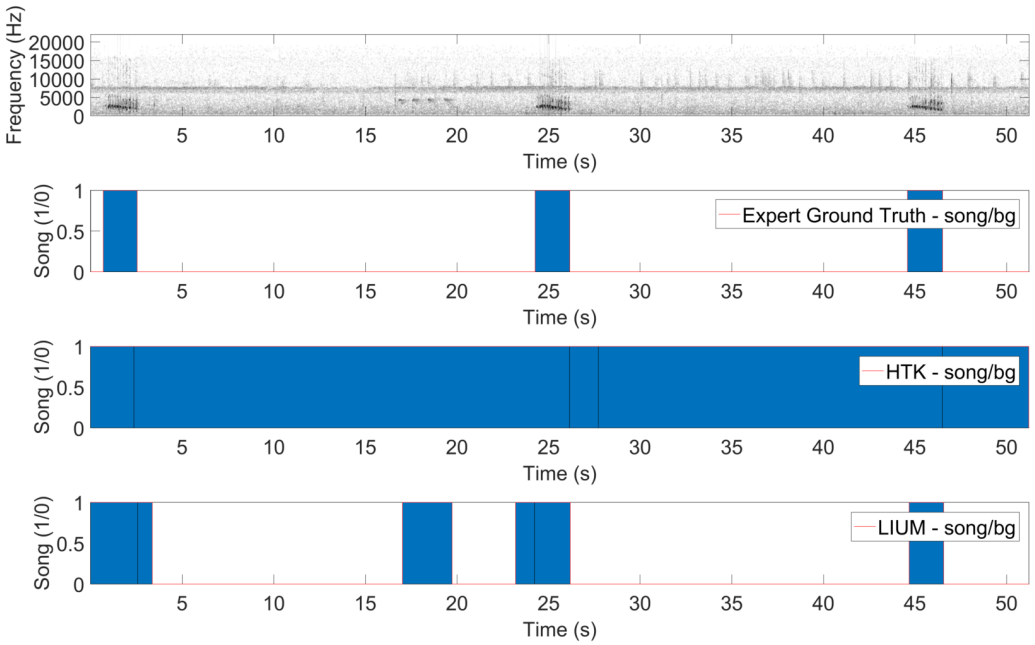


Figure 6.4: Herpsilochmus stictocephalus recording spectrogram (top subplot), ground truth annotation by expert listener (second subplot), HTK HMMs segmentation result (third subplot), LIUM segmentation result (bottom subplot). Blue area corresponds to song label, white corresponds to background or non-target label. For this example, HTK *PF* is 88.93% and for LIUM *PF* is 10.71%.

song observed. HMMs perhaps accidentally get this segmentation correct because it is unable to identify the non-target bird song in the noisy example.

Sounds produced by insects are commonly observed in many bird recordings. Insects sounds tend to be continuous, i.e. frequency remains stationary and duration spans the whole recording. LIUM deals with these types of non-target audio quite well. Insect sounds appeared to have little effect on boundary selection in experiments, as the sound was continuous. HMMs however treat a limited set of adjacent frames, dependent on the number of HMM states, separately and compares the frames to song and background models. Insect sounds tend to be narrowband and periodic and were often falsely identified as song using HMMs where the background model is trained on a mixture of data, with and without insects.

LIUM is better at rejecting low amplitude non-target song, whereas the HMM system is susceptible to false alarm of quiet non-target song. Some recordings have recordist speech at the beginning and end of a recording where metadata is described. Surprisingly speech was identified as target song in certain instances, even though speech was included in training data for the non-target model, whereas no speech was present in the target song training data. There were also cases in the data with overlap of target and non-target song. HMMs generally overestimated

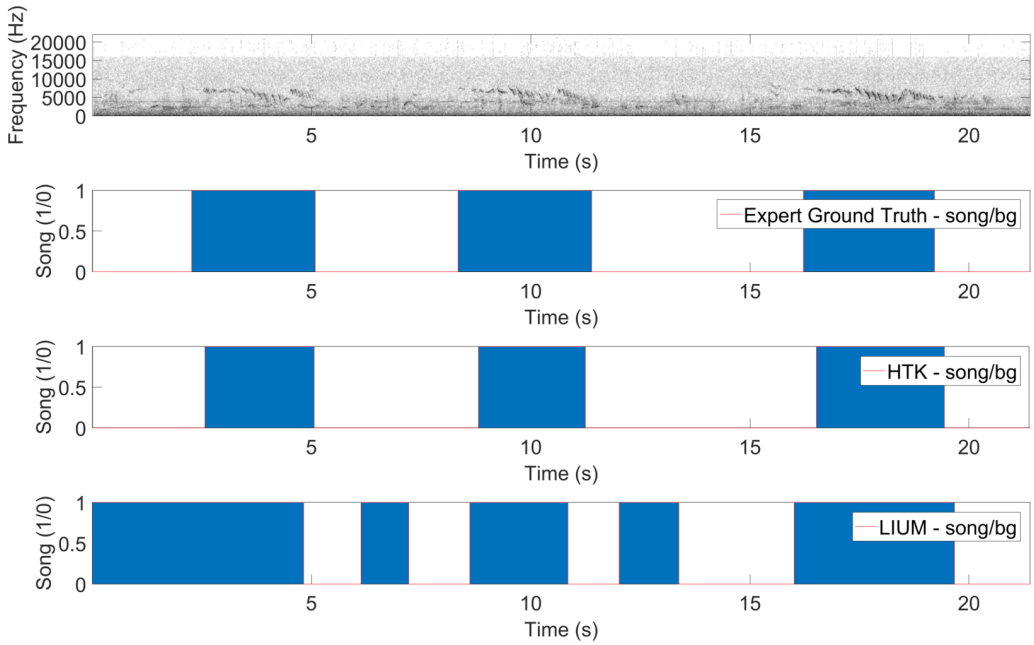


Figure 6.5: Certhia familiaris recording spectrogram (top subplot), ground truth annotation by expert listener (second subplot), HTK segmentation result (third subplot), LIUM segmentation result (bottom subplot). Blue area corresponds to song label, white corresponds to background or non-target label. For this example, HTK *PF* was 6.84% and for LIUM *PF* was 30.21%.

the length of song bouts in these cases, whereas LIUM was more robust and identified song boundaries with a better accuracy.

6.4.2 Extension of Diarization and Clustering

There were two recordings which contained two individuals of the same species vocalizing back and fourth. When one would sing the other would respond. These recordings also contained background chorus of non-target birds. For a Cercomacra Serva recording containing two birds in the foreground both of same target species, the LIUM system identified 3 target clusters but each of the 3 clusters contained predominately song instances from target bird 1. An extract of that file ground truth annotation and LIUM segmentation clustering is plotted in Figure 6.6. The top subplot shows ground truth annotations for the first 24 seconds. Yellow corresponds to song from bird 1 and green corresponds to song from bird 2. The bottom subplot shows LIUM segmentation results before clusters are merged into a single song class label. Purple corresponds to cluster 1 and red corresponds to cluster 2. The results show that LIUM identifies bird 1 as the target and rejects bird 2. Variation in the song of bird 1 causes LIUM to assign more than a single cluster. Because of the second bird LIUM outputs incorrect boundaries of bird 1's song,

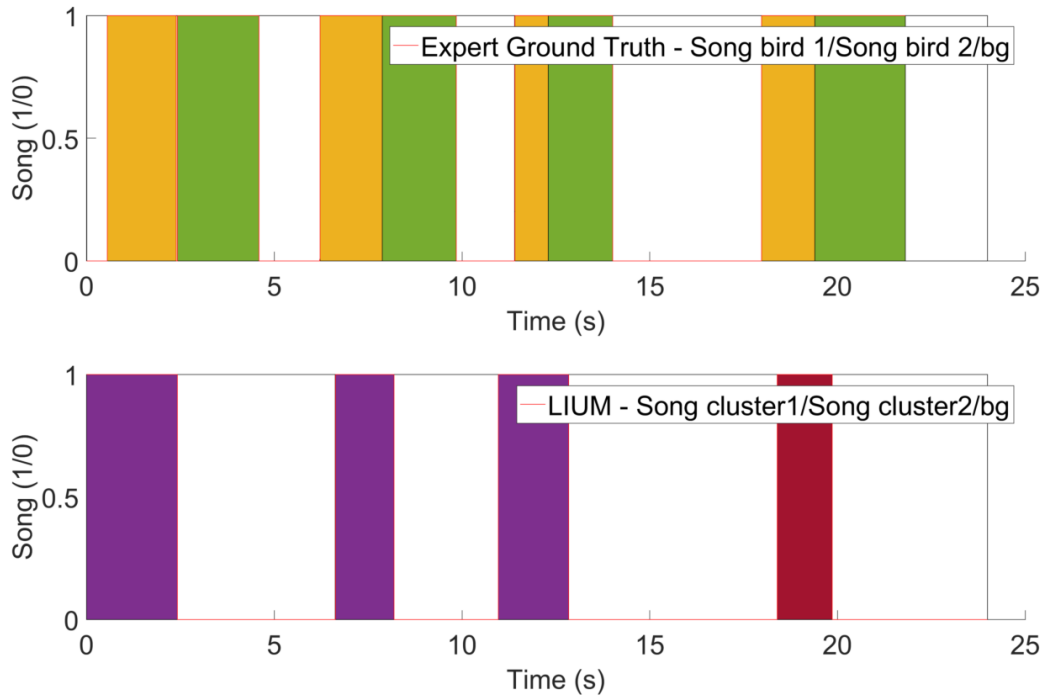


Figure 6.6: Cercomacra Serva recording containing two birds in the foreground both of same target species. Top subplot shows ground truth annotations for the first 24 seconds. Blue corresponds to song from bird 1 and green corresponds to song from bird 2. The bottom subplot shows LIUM segmentation results before clusters are merged into a single song class label. Purple corresponds to cluster 1 and red corresponds to cluster 2.

confusing it with bird 2 at 12-13 seconds in Figure 6.6.

The result shown in Figure 6.6 suggests LIUM clusters variability in the song of target bird 1 rather than identifying different targets as different clusters. With some modifications to the system perhaps a full diarization approach for birds is possible, where "who sang when" annotations could be made. Some normalization technique which takes into account how far away a target is from the recorder may need to be included, as the approach used here rejects non-target bird song which sounds faint or comes from the background. The current system successfully segments a recording with a single target. Perhaps with some modifications there is potential to use speaker diarization techniques to segment recordings with multiple birds present with start and end times of all syllables and a label for what source each syllable came from. An extension to this would be automatic counting of species or counting the number of individuals of the same species in a recording. Another application could be studying variability of song within a species, i.e. the greater the variability the greater number of clusters assigned to a recordings of a species.

Figure 6.7 shows a scatter plot which compares the number of clusters output by LIUM

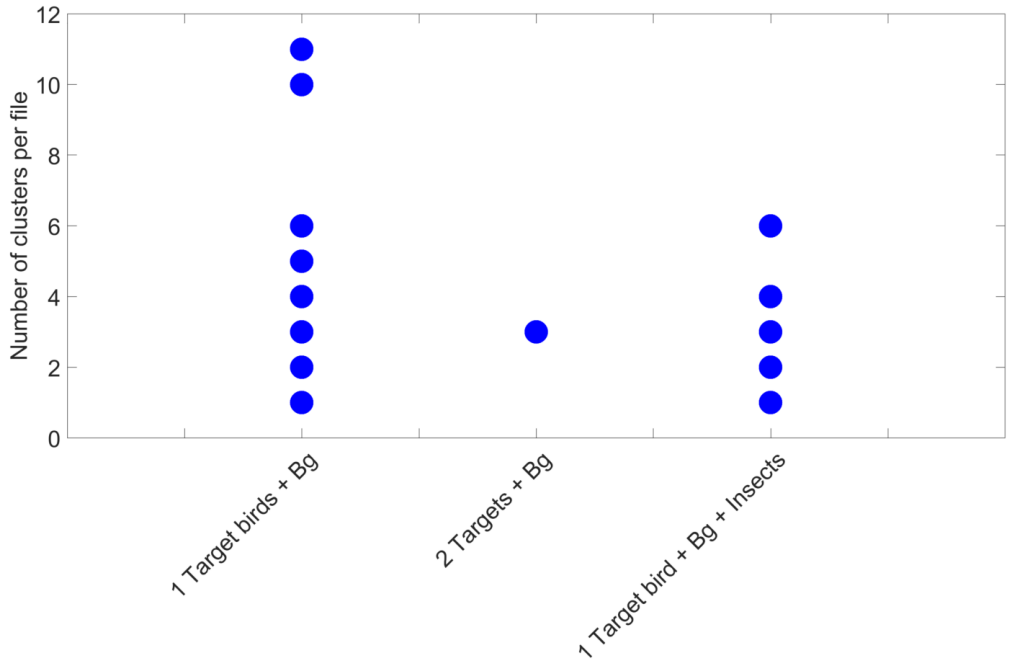


Figure 6.7: Scattter plot showing the number of target song clusters assigned by LIUM when input file has 1 target bird plus background sounds, 2 target birds plus background sounds and 1 target bird plus background sounds plus insect sounds. This plot demonstrates number of clusters is not overly reliant on these factors.

when files contain one target bird plus background sounds, two target birds plus background sounds and one target bird plus background sounds plus insect sounds. No trend emerges here, as it is more likely variability in target bird's vocalizations is the main factor in the number of clusters assigned. It would be interesting to see if whistles, trills, nasal sounds, noisy and polyphonic sounds were each assigned different clusters within the same species, i.e. syllable level classification. Earlier in Section 6.2.4, it was stated that multiple clusters were grouped together as song, since the application here was target/non-target delimitation. Initial work on this approach could explore the potential to further cluster results into different elements of song.

Segmentation accuracy could be improved upon with further investigation into feature extraction methods. Perhaps using some sort of location inference would help identify where birds are in space. Rejecting sounds coming from further away would improve segmentation accuracy of individuals near the recorder's microphone.

6.5 Conclusion

This work demonstrates that a speaker diarization approach to bird song segmentation can improve on HMM-based segmentation. It satisfies the need for a general purpose segmentation approach that generalizes well to a variety of situations, recording conditions and species. HMM based segmentation is limited to a range of data already seen. LIUM assesses each recording through hierarchical clustering and change point detection. LIUM is more robust to changes in noise composition across recordings and therefore offers a better segmentation method.

Further potential refinements include the use of alternative feature sets and refining the segmentation to label different elements within a song bout of a target bird.

7 Conclusion

This thesis presented "Digital Signal Processing Approaches to Bird Song Analysis". The major themes explored in this thesis have been measuring the difference of vocalizations across bird populations, the accurate extraction of pitch in bird vocalizations, and the segmentation of vocalizations of longer field recordings. This chapter now revisits the major findings from each chapter in the thesis, considering where future directions for this work may lie.

7.1 Chapter 2: A review of bird acoustic analysis

Chapter 2, provided background on bird song analysis used in subsequent Chapters. An overview of bird vocalizations was given and their importance stated. Characteristics of vocalizations were described in detail. Automatic acoustic analysis by engineers was reviewed. Previous work on bird species classification was examined and relevant techniques described.

7.2 Chapter 3: Quantifying difference in vocalizations of bird populations

In chapter 3 a measure of difference for bird calls and songs based on the difference in distribution of pitch contours was presented. The difference measure was inspired by dialect distance from speech processing. A novel Vector Quantization (VQ) codebook was implemented which clusters pitch contours. The addition of VQ codebooks produced results similar to Zoologists' opinion of difference for a set of bird populations. Experiments were also preformed on reduced datasets

and subsets of the same population to evaluate performance. The reported experiments give confidence that a system can be developed that quantifies difference of bird vocalizations between populations. With further research into difference threshold, regions can be determined to aid ornithologists in the species/subspecies decision making process.

7.3 Chapter 4: YIN-bird: Improved pitch tracking for birds

Work in chapter 4 presented an improved Pitch Detection Algorithm (PDA) implementation for bird song. Bird sound analysis using signal processing and machine learning techniques is in its early stages. Pitch is not only important for analysis and synthesis, but is used in measuring bird population similarity. This relies on accurate pitch estimation. Bird frequency range varies dramatically from species to species, and even within syllables in a song repertoire from a single bird. Hence static YIN parameters are not useful in bird recordings. Automatically determining the $f_{0_{min}}$ parameter on a segment by segment basis for YIN improves pitch estimation. This improvement should in turn improve accuracy on bird species and phrase comparisons, allowing fully automatic batch processing of large numbers of recordings from different species.

A qualitative description of a range of vocalizations that pose problems for PDAs was also presented in detail and examples shown. Knowledge of when to use YIN-bird, depending on what type of vocalizations are being analyzed, is informative for zoologists. Pitch estimates of whistles, harmonics with strong Fundamental Frequency ($\boldsymbol{F_0}$) and trills should be treated with confidence while pitch tracking on two-voiced and nasal sounds remains problematic. Investigation into pitch analysis of these types of vocalizations would be interesting future work.

The value of YIN-bird lies not only in this demonstrated performance improvement. The possibility of fully automatic processing of bird song to extract pitch will allow researchers to process larger amounts of data, improving certainty in tasks such as species comparison based on pitch measurements.

7.4 Chapter 5: Species delimitation inspired by zoological techniques

Work in chapter 5 presented automatic means of extracting features from bird song recordings for a standardized criteria of bird species delimitation. Pairs of birds were compared to help quantify what level of vocal difference is required to classify two groups as different species. The reclassification of bird species is a multifaceted task which requires investigation of morphological, visual, acoustic and behavioral characteristics. Work in chapter 5 concentrated on acoustic analysis to aid ornithologists with their work. Current practice involves manual inspection of spectrograms which is tedious and subjective. YIN-bird was utilized to extract spectral features from recordings of bird song. Features automatically extracted from song (together with temporal features extracted using manual segmentation) gave vocal scores of similar proportion

compared to the scores presented by Tobias et al. in [160]. A comparison between Tobias et al.'s vocal score and the bird difference metric presented in chapter 3 was also discussed. Difference scores from both systems for some pairs agreed while other pairs did not. The inconsistencies were explained due to the nature by which each system measures difference. Tobias et al.'s method looks at temporal features and frequency features at a high level while the bird difference metric studies finer pitch contour structures. Songs with similar duration and bandwidth will give a low difference score using Tobias et al.'s method, whereas the bird difference measure will give a high difference score, if the pitch contour structures of these songs are very different. The bird difference metric however does not capture important information like song duration. Both systems have advantages and disadvantages and a combination of both systems may give ornithologists the clearest picture in their evaluations. Future work involves automatic temporal feature extraction which would deal with the difficult task of estimating number of notes per song automatically for bird recordings.

7.5 Chapter 6: Segmentation of Field Recordings

Finally, work in chapter 6 presented automatic segmentation of bird song recordings using speaker diarization. This work demonstrated that a speaker diarization approach to bird song segmentation can improve on Hidden Markov Model (HMM)-based segmentation. It satisfies the need for a general purpose segmentation approach that generalizes well to a variety of situations, recording conditions and species. Further potential refinements include the use of alternative feature sets and refining the segmentation to label different elements within a song bout of a target bird.

7.6 Final Remark

It is hoped that the findings of this thesis will inspire the work of others in the field of bird song analysis and digital signal processing in general. As research into bird taxa reclassification matures in the field of zoology, advances in digital signal processing will continue to aid ornithologists with their research. As datasets of bird recordings continue to expand, it will become more of a big data issue where advances in machine learning will help automate the analysis process. Some of the current problems involve classification of birds in noisy environments, pitch extraction when multiple targets are present, automatic population counting, identifying individuals within a species, is there a bird present in a recording, and how many individuals need to be sampled to compare two populations? Future work could involve obtaining recordings of birds without humans in the loop, creating a fully automated bird monitoring and classification system.

A

Species delimitation data

Table A.1: Description of each bird population's calls and song, as described by expert listener. This data is a subset of data used by Tobias et al in [160] and was used for experiments in Chapter 5. Writing in *italics* signifies expert listener's phonetic description of bird vocalization.

Pair No.	**Genus**	**Species**	**Song Description**	**Call Description**
5	Arremon	brunneinuchus	≈ 5-6 very high pitched squeaky, sibilant, slurred whistles, that alternate in slurr direction. *pee-tee-zee-pee-tzee-teezee*	Thin, short and sibilant *piik.*
	Arremon	torquatus	≈ 3 very high pitched squeaky, sibilant whistles, often steeply slurred (often alternating in slurr direction). *pee-tzee-pee* or *pee-tzuch-pee-tzee*	Thin, short, sibilant and sharply downslurred *pwik.*
6	Cercomacra	nigrescens	Single low overslurred *wip/wob* followed by a higher pitched rattling trill, variable in speed, and downslurred. *wob-wibwibwibwib*	Rattly low pitched, nasal and noisy notes (almost crow like) *grreah.*

Table A.1 - Continued on next page.

Table A.1 - Continued from previous page.

	Cercomacra	serva	≈ 6-7 short, often overslurred whistles, increasing both in speed and pitch, with a sudden end. The first note is the strongest and longest, and there tends to be a pause between it and the rest of the song. *wub–wib-wib-wib-wib-buib*	Rattly, harsh, unmusical, low pitched trill of noisy notes *grrrahgragragragra* - variable length.
7	Certhia	brachydactyla	≈ 5-6 high pitched slurred whistles, evenly spaced with 2-3 string of downslurred notes at the beginning, and a general upslurred trend for second half of the series, with a flourished ending. *teet-teet-ee-roi-tee-tiit*	Pleasant, repeated, high-pitched short whistle, *tyt-tyt-tyt-tyt* or *piit-piit-piit.*
	Certhia	familiaris	Quick and high pitched song that is somewhat variable yet still made of distinct phrases, and with a general downslurred trend. First 4 high pitched overslurred/downslurred whistles, followed by 1 or 2 downslurred trills, sometimes interspersed with sharp downslurred whistles, and finished with a sharp underslurred/upslurred flourish, *tswii-tswii-tswii-tritritri-ts–tsi-tritritritri-twoo-ih*	A couple of call types - a thin and high pitched *sit*, and most distinctively a rich burry *tree*, that has a vibrato quality, and is often repeated in sequence, sounding like a cricket call.
10	Criniger	calurus	≈ 4-5 notes, evenly spaced, two short *pips* followed by two very slurred (wavy) notes, *pip-pip-weoh-weoh*, overall somewhat noisy and nasal quality.	Short repeated overslurred whistles, *wiiouh-wiiouh.*

Table A.1 - Continued on next page.

Table A.1 - Continued from previous page.

	Criniger	ndussumensis	≈ 3 burry notes, one short, one long, and one longer, with the first variable in slurr direction and the last two increasing in downslurr. *pu-perr-peorr*	No calls in data.
12	Empidonax	hammondi	A 3 part well spaced song, with three distinct phrases. First, an abrupt two syllable phrase, followed by a low burry note, and finished with another two syllable phrase of a sharply upslurred whistle and a short upslurred burr. *ssilit-greep-sliipeet*	Call a sharp, short, overslurred whistle, *pip*, *peep*, or *peek*.
	Empidonax	oberholseri	Also a 3 part song, first a vaguely two syllabled *prll-it*, increasing in pitch, then a rough burry low pitched *prrdrrt*, and finished with a clear high pitched and overall upslurred *pseet*. *prll-it-prrdrrt-pseet*	Most commonly a dry, soft, *whit*, similar to a Willow Flycatcher. Occasionally the males will give another call, a melancholy *dew-hic*.
19	Herpsilochmus	stictocephalus	A downslurred "laugh" series of dry notes (≈ 12-15), relatively long, which briefly accelerates before a long downslurred de-acceleration.	Short, sharp *wit*.
	Herpsilochmus	sticturus	Bouncing ball trill, ≈ 5 slow initial strong bounce notes, then acceleration, *ch-ch-ch-ch-ch-chchchchch*. Monotonous & lower pitch, and longer than the stictocephalus.	Again, a short sharp *wit* or *chit*.

Table A.1 - Continued on next page.

Table A.1 - Continued from previous page.

41	Picoides	pubescens	≈13-15 rapidly repeated and high pitched nasal notes. Tends to accelerate and then slightly downslurr towards the end of the series, but overall the speed can be variable. Sometimes there is a bouncing ball quality, and can sound somewhat sibilant from the high pitch, *sii-sii-sii-sii-sii-sisisisisisisi* or *piit-piit-piit-piit-pii-piipiipiipiipipipipi*	No calls in data.
	Picoides	villosus	Similar to pubescens but overall lower pitched. A nasal *keek* is repeated, but this varies in number within a series. Less nasal, *keek-keek-kee-kiikiikiikiikii*, but does not descent in pitch at the end.	Short, sharp peek note very similar to pubescens, but louder and slightly lower pitched and often sounding more emphatic.
44	Regulus	ignicapillus	Very high pitched series of whistled, sibilant notes, ascending in frequency and delivered in a crescendo (upslurred).	Very high pitched *zit-zit-zit-zit.*
	Regulus	regulus	Characteristic sqeaky bicycle wheel song, extremely high pitched & sibilant. Variable *sii-pa-pa* is repeated 5-7 times, and ends in a variable flourish. *sii-pa-pa-stichi-see-pee*	Even higher pitched and sibilant, less rough than the ignicapillus, *zii-zii-zii-zii.*

Table A.1 - Continued on next page.

Table A.1 - Continued from previous page.

45	Seiurus	motacilla	Series of shrill and slightly nasal slurred whistles, slightly descending in pitch with the first 3-4 notes strong and evenly spaced and followed by a flourish of rapid whistles alternating in slurr direction and ending with a sharp overslurr, and sometimes final short whistle. *peeoh-peeoh-peeoo-poohpo -chi-chi-pweeoh-pi*	A harsh high and a bit noisy *chik*.
	Seiurus	noveboracensis	Again, a series of 3-4 slurred whistles followed by a flourish, however, the song is shorter than the motacilla, and the flourish is not as varied, and has greater bandwidth. *wip-wip-wip-wip- weoh-weoh-wiii-wi-weoii*	Call a loud, hard, short *spwiK* with a strong *K*sound.
53	Thamnophilus	murinus	Low pitched and nasal, ≈2-4 repeated nasal notes, wit the last note a sharp nasal overlurr (two syllable). *neah-neah-neah-nii-mah*	nasal *cah*
	Thamnophilus	schistaceus	Similar to murinus, but higher pitched, faster, and with more repeated nasal notes before the overslurred end note, *muh-muh-muh-muh- muh-mehnuh*. Slight acceleration before the last note.	Nasal and slightly burry *rreah*.

Table A.1 - Continued on next page.

Table A.1 - Continued from previous page.

54	Toxostoma	bendirei	Very complex!! Clear, loud and melodious warble with some repetitive segments, and continuing at length. Highly variable and rich, includes nasal notes, burrs, buzzes, polyphony, whistles, chirps etc. essentially everything!	Soft, throaty chuck.
	Toxostoma	curvirostre	Simpler and shorter than the bendirei, with distinct pauses, and more whistles than burrs, and other types of notes. Long series of varied phrases, a pleasant and melodious warble.	Sharp whistled *whit-wheet.*
55	Tyrannus	verticalis	Rising series of sharp nasal *kips*, culminating in a frantic burst of loud descending notes, *kip-kip-kuh-wee-wihwihwih.* Squeaky quality. First number of *kips* are variable, it's the fast and rapidly descending *kuh-wee-wihwihwih* that is most constant between songs.	Sharp *wit or kit* .

Table A.1 - Continued on next page.

Table A.1 - Continued from previous page.

	Tyrannus	vociferans	Rapid rasping nasal song, complex, first a *chi-bEW-chi-beer*, made of a short chirp and a slurred nasal note, and then repeated. This two syllable phrase can vary in number. Then follows a complex and rapid *chew-bu-bii-ber-bih-bi*, made of burry notes and nasal notes of varying direction, giving it a saw like quality.	Nasal *chi-bew.*
57	Vermivora	chrysoptera	≈ 4-5 buzzy notes, the first long and somewhat drawn out, then 3-4 lower pitched and repetitive shorter, monotonous buzzy notes. *bEEzt-bzz-bzz-bzz.*	High pitched lightly buzzy chip or *tzip.*
	Vermivora	pinus	Long buzzy note followed by a noisy trill. The buzzy note is sometimes underslurred/slurred, and the trill is usually higher pitched than the buzzy note (but not always). *bzeee-btzbtzbtzbtzbtz*	A harsh high chip.
58	Vireo	olivaceus	Rapid, usually two syllabled throaty whistle. Variable, but the two whistles tend to be musical, very slurred (wavy), with the first whistle usually higher pitched, *pee-oo-weeh*, or *pee-owoh.*	Rough, burry, noisy and somewhat nasal *mrreah.*

Table A.1 - Continued on next page.

Table A.1 - Continued from previous page.

Vireo	philadelphicus	Variable, and similar to the olivaceus, but with usually more whistles, and sometimes relatively complex with ≈ 5 notes, and some burry short and nasal notes, all in a rapid downslurr series.	Noisy *rreh*.

End of Table A.1.

B

Hidden Markov Models (HMMs) theory

This Appendix contains detailed description of Hidden Markov Models (HMMs).

Rabiner [126] describes three problems for HMMs. Firstly the probability of an observation sequence given a particular HMM needs to be computed. Secondly, given an observation sequence, what the most likely state sequence to produce that observation sequence is, needs to figured out. And thirdly, given some training observations, the model parameters must be adjusted. An observation sequence X (a feature vector extracted from a test recording in the case of speech recognition) contains a vector of n-dimensional features, $X = (x_1, x_2, ..., x_T)$, where T is the number of frames in the feature vector. x_t is an n-dimensional vector of a single frame. Each frame's state in a sequence is denoted by q_t, where t is the frame number, and the whole state sequence is denoted by Q. Computing $P(X|\lambda)$ solves the first problem. Given feature vector X there are a number of possible sequences of T states. Any particular state sequence Q is given in Eq. B.1. The probability of X given Q, coming from model λ is calculated like a Gaussian Mixture Model (GMM) for each state in Eq. B.2. The probability of a particular state sequence is given in Eq. B.3. The joint probability of a given feature sequence with a certain state sequence, i.e. the probability that X and Q occur simultaneously (given model λ) is given by Eq. B.4.

$$Q = q_1 \; q_2 \; ... \; q_T \tag{B.1}$$

$$P(X|Q,\lambda) = \prod_{t=1}^{T} P(x_t|q_t,\lambda) = b_{q_1}(x_1) \; b_{q_2}(x_2) \; ... \; b_{q_T}(x_T) \tag{B.2}$$

$$P(Q|\lambda) = \pi_{q_1}\ a_{q_1 q_2}\ a_{q_2 q_3}\ ...\ a_{q_{T\text{-}1}\, q_T} \tag{B.3}$$

$$P(X, Q|\lambda) = P(X|Q, \lambda)\ P(Q, \lambda) \tag{B.4}$$

The overall probability of X given the model sums over all possible state sequences of length T. This is given in Eq. B.5. To sum over all possible state sequences is computationally expensive. In practice it is not used, instead the forward-backward procedure is employed [126].

$$P(X|\lambda) = \sum_{ALL\ Q} P(X|Q, \lambda)\ P(Q, \lambda) \tag{B.5}$$

$$= \sum_{q_1 q_2 ... q_T} \pi_{q_1} b_{q_1}(x_1)\ a_{q_1 q_2} b_{q_2}(x_2)\ a_{q_2 q_3} b_{q_3}(x_3)\ ...\ a_{q_{T\text{-}1} q_T} b_{q_T}(x_T)$$

The second problem, computing the most likely state sequence, uses the Viterbi algorithm. This outputs an optimum state sequence, $q_1 q_2 ... q_T$, given feature vector $X = (x_1,\ x_2,\ ...,\ x_T)$. Eq. B.6 defines $\delta_t(i)$ the best score along a single path, at time t, that accounts for the first t observations and ends in state S_i. As shown in Figure B.1, this algorithm can be visualized as finding the best path through a matrix where the vertical dimension represents the states of the HMM and the horizontal dimension represents the frames of speech (i.e. time). Each large dot in the picture represents the log probability of observing that frame at that time and each arc between dots corresponds to a log transition probability [175].

$$\delta_t(i) = \max_{q_1 q_2 ... q_{t-1}} P[q_1 q_2\ ...\ q_t = i,\quad x_1 x_2\ ...\ x_t|\lambda] \tag{B.6}$$

Then by induction Eq. B.7 gives best path to time $t+1$ ending in state j.

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] b_j(x_{t+1}) \tag{B.7}$$

To retrieve the state sequence, array $\psi_t(j)$ is used to keep track of the states at each time instance. The complete procedure for finding the best state sequence starts with Eq. B.8 and Eq. B.9.

$$\delta_1(i) = \pi_1\ b_i(x_1) \tag{B.8}$$

$$\psi_1(i) = 0 \tag{B.9}$$

This is followed by recursively selecting the best state sequence for each frame in Eq. B.10 and Eq. B.11.

$$\delta_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}]b_j(x_t) \qquad 2 \le t \le T \qquad 1 \le j \le N \tag{B.10}$$

$$\psi_t(j) = \max_{1\le i\le N}[\delta_{t-1}(i)a_{ij}] \qquad 2 \le t \le T \qquad 1 \le j \le N \tag{B.11}$$

The procedure terminates with Eq. B.12 and Eq. B.13.

$$P^* = \max_{1\le i\le N}[\delta_T(i)] \tag{B.12}$$

$$q_T^* = \arg\max_{1\le i\le N}[\delta_T(i)] \tag{B.13}$$

Eq. B.14 is path (state sequence) backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, ..., 1. \tag{B.14}$$

The third problem for HMMs is determining a method to adjust the model parameters (A, B, π) to maximize the probability of the observation sequence given the model [126]. During training the model parameters are locally maximized using an iterative process such as Baum-Welch method. Given any finite observation sequence as training data, there is no optimal way of estimating the model parameters [126], therefore $P(X|\lambda)$ is locally maximized iteratively.

$\xi_t(i,j)$, the probability of being in state S_i at time t, and state S_j at time t+1, given the model and observation sequence is defined in Eq. B.15.

$$\xi_t(i,j) = P(q_t = S_i,\ q_{t+1} = S_j|X, \lambda) \tag{B.15}$$

A backward variable $\beta_t(i)$ is defined in Eq. B.16.

$$\beta_t(i) = P(x_{t+1}\ x_{t+2}\ ...\ x_T|q_t, \lambda) \tag{B.16}$$

In terms of forward and backward variables ξ_t is expressed in another form in Eq. B.17.

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(x_{t+1})\ \beta_{t+1}(j)}{P(X|\lambda)} = \frac{\alpha_t(i)a_{ij}b_j(x_{t+1})\ \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)\ a_{ij}b_j(x_{t+1})\ \beta_{t+1}(j)} \tag{B.17}$$

where $\alpha_t(i)$, a forward variable, is the probability of a partial observation sequence until time t, given the model λ.

γ_t is the probability of being in state S_i at time t, given observation sequence and the model (see Eq. B.18).

$$\gamma_t(i) = P(q_t = S_i|X, \gamma) \tag{B.18}$$

The relation between $\gamma_t(i)$ and $\xi_t(i,j)$ is given in Eq. B.19.

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{B.19}$$

$\sum_{t=1}^{T-1} \gamma_t(i)$ is the number of expected transitions from S_i and $\sum_{t=1}^{T-1} \xi_t(i,j)$ is the expected number of transitions from S_i to S_j. The parameters of the HMM are iteratively reestimated using Eq. B.20, Eq. B.21 and Eq. B.22.

$$\bar{\pi}_i = \gamma_1(i) \tag{B.20}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{B.21}$$

$$\bar{b}_j(k) = \frac{\sum_{\substack{t=1 \\ s.t. x_t = v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{B.22}$$

The iterations are profitable since Baum proved, that $P(X|\bar{\lambda} \geq P(X|\lambda)$ is optimized until they reach a limiting point. The result is called Maximun Likelihood (ML) estimate of the HMM. The algorithm will reach local maximum if initialization is good. The reestimation formulas, Eq. B.20, Eq. B.21 and Eq. B.22, can be derived by maximizing Baum's auxiliary function [126], Eq. B.23.

$$Q(\lambda, \bar{\lambda}) = \sum_{Q} P(Q|X, \lambda) \log P(X, Q|\bar{\lambda}) \tag{B.23}$$

The Baum-Welch algorithm is similar to the Expectation Maximization (EM) algorithm of statistics, in which the expectation step corresponds to the calculation of the auxiliary function and the modification step to the reestimation of λ.

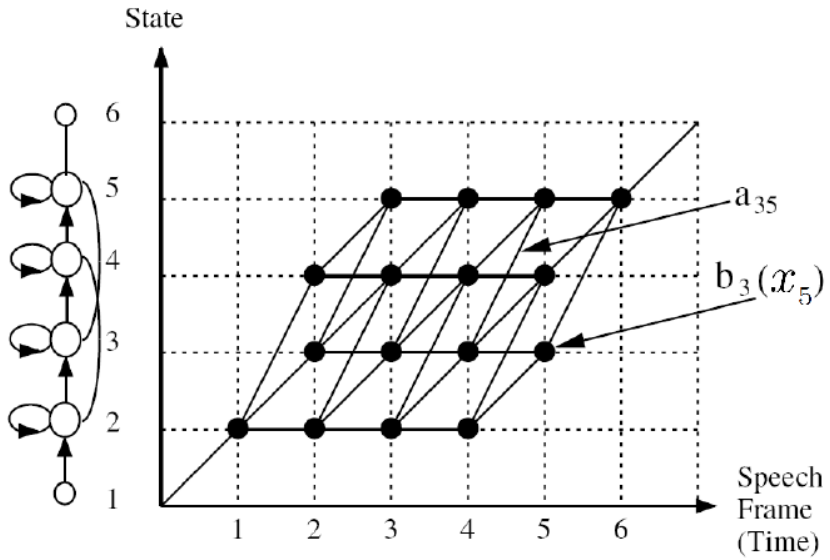


Figure B.1: The Viterbi algorithm for isolated word recognition [175].

Bibliography

[1] I. Agranat. Automatic detection of cerulean warblers using autonomous recording units and song scope bioacoustics software. *Wildlife Acoustics, Concord, Massachusetts, USA*, 2007. http://www.wildlifeacoustics.com/ (accessed 07 Dec 2016).

[2] I. Agranat. Automatically identifying animal species from their vocalizations. In *Fifth International Conference on Bio-Acoustics, Holywell Park*, 2009.

[3] P. Alstrom and R. Ranfft. The use of sounds in avian systematics and the importance of bird sound archives. *BULLETIN-BRITISH ORNITHOLOGISTS CLUB*, 123:114–135, 2003.

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

[5] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit. A comparative study of pitch extraction algorithms on a large variety of singing sounds. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7815–7819, 2013.

[6] P. C. Bagshaw, S. M. Hiller, and M. A. Jack. Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pages 1003–1006, 1993.

[7] G. J. Beckers. Bird speech perception and vocal production: a comparison with humans. *Human biology*, 83(2):191–212, 2011.

[8] G. J. Beckers, R. A. Suthers, and C. Ten Cate. Pure-tone birdsong by resonance filtering of harmonic overtones. *Proceedings of the National Academy of Sciences*, 100(12):7372–7376, 2003.

[9] E. Benetos and S. Dixon. Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. *The Journal of the Acoustical Society of America*, 133(3):1727–1741, 2013.

[10] E. Benetos, M. Lagrange, and M. D. Plumbley. Detection of overlapping acoustic events using a temporally-constrained probabilistic model. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 6450–6454. IEEE, 2016.

[11] F. Beritelli, S. Casale, and G. Ruggeri. Performance evaluation and comparison of ITU-T/ETSI voice activity detectors. In *Acoustics, Speech and Signal Processing (ICASSP), 2001 IEEE International Conference on*, volume 3, pages III–1425. IEEE, 2001.

[12] F. Biadsy and J. B. Hirschberg. Using prosody and phonotactics in arabic dialect identification. In *INTERSPEECH 2009 – 10th Annual Conference of the International Speech Communication Association, September 6–10, Brighton, UK*, pages 208–211, 2009.

[13] J. A. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126, 1998.

[14] D. T. Blumstein and O. Munos. Individual, age and sex-specific information is contained in yellow-bellied marmot alarm calls. *Animal Behaviour*, 69(2):353–361, 2005.

[15] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pages 97–110. Amsterdam, 1993.

[16] P. Boersma and D. Weenink. Praat: Doing phonetics by computer (version 5.2.22), 2010. http://www.fon.hum.uva.nl/praat/ (accessed 29 Nov 2016).

[17] A. L. Borker, M. W. McKown, J. T. Ackerman, C. A. Eagles-Smith, B. R. Tershy, and D. A. Croll. Vocal activity as a low cost and scalable index of seabird colony size. *Conservation Biology*, 28(4):1100–1108, 2014.

[18] T. S. Brandes. Feature vector selection and use with hidden markov models to identify frequency-modulated bioacoustic signals amidst noise. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1173–1180, 2008.

[19] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. 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.

[20] M. Brookes. Voicebox: Speech processing toolbox for matlab, 2005. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (accessed 15 Apr 2016).

[21] R. T. Buxton and I. 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.

[22] J. Cai, D. Ee, B. Pham, P. Roe, and J. Zhang. Sensor network for the monitoring of ecosystem: Bird species recognition. In *Intelligent Sensors, Sensor Networks and Information, 2007. ISSNIP 2007. 3rd International Conference on*, pages 293–298. IEEE, 2007.

[23] A. Camacho. *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*. Ph.d. dissertation, university of florida, usa, 2007.

[24] A. Camacho and J. G. Harris. A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, 124(3):1638–1652, 2008.

[25] C. K. Catchpole and P. J. Slater. *Bird song: biological themes and variations, 2nd Edition*. Cambridge University Press, ISBN 9780521872423, 2008.

[26] R. Charif, D. Ponirakis, and T. Krein. Raven bioacoustic software - cornell laboratory of ornithology, ithaca, ny, 2014. http://www.birds.cornell.edu/brp/raven/RavenOverview.html (accessed 07 Dec 2016).

[27] R. Charif, D. Ponirakis, and T. Krein. Xbat - bioacoustic software for matlab, 2014. https://code.google.com/archive/p/xbat-devel/ (accessed 07 Dec 2016).

[28] Z. Chen and R. 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.

[29] D. Chesmore. Automated bioacoustic identification of species. *Anais da Academia Brasileira de Ciências*, 76(2):436–440, 2004.

[30] L. Chi-Yueh and H. C. Wang. Language identification using pitch contour information in the ergodic markov model. In *Acoustics, Speech and Signal Processing (ICASSP), 2006 IEEE International Conference on*, volume 1, pages I–I, 2006.

[31] W. Chu. *Noise Robust Signal Processing for Human Pitch Tracking and Bird Song Classification and Detection*. Ph.d. dissertation, university of california, los angeles, usa, 2012.

[32] W. Chu and A. Alwan. A correlation-maximization denoising filter used as an enhancement frontend for noise robust bird call classification. In *INTERSPEECH 2009 – 10th Annual Conference of the International Speech Communication Association, September 6–10, Brighton, UK*, pages 2831–2834, 2009.

[33] P. J. Clemins, M. T. Johnson, K. M. Leong, and A. Savage. Automatic classification and speaker identification of african elephant (loxodonta africana) vocalizations. *The Journal of the Acoustical Society of America*, 117(2):956–963, 2005.

[34] J. Cohen. Statistical power analysis for the behavioral sciences . hilsdale. *NJ: Lawrence Earlbaum Associates*, 2, 1988.

[35] N. Collar. Taxonomy and conservation: chicken and egg. *BULLETIN-BRITISH ORNITHOLOGISTS CLUB*, 117:122–136, 1997.

[36] N. Collar. A partial revision of the asian babblers (timaliidae). *Forktail*, 22:85, 2006.

[37] S. Collins. Vocal fighting and flirting: the functions of birdsong. *Nature's music: the science of birdsong*, pages 39–79, 2004.

[38] E. F. Connor, S. Li, and S. Li. Automating identification of avian vocalizations using time-frequency information extracted from the gabor transform. *The Journal of the Acoustical Society of America*, 132(1):507–517, 2012.

[39] Cornell. Cornell laboratory of ornithology, ithaca, new york, usa, 2015. http://www.birds.cornell.edu/Page.aspx?pid=1478 (accessed 10 May 2016).

[40] C. V. Cotton and D. P. Ellis. Spectral vs. spectro-temporal features for acoustic event detection. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 69–72. IEEE, 2011.

[41] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

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

[43] K. De Queiroz. Ernst mayr and the modern concept of species. *Proceedings of the National Academy of Sciences*, 102(suppl 1):6600–6607, 2005.

[44] V. B. Deecke and V. M. Janik. Automated categorization of bioacoustic signals: avoiding perceptual pitfalls. *The Journal of the Acoustical Society of America*, 119(1):645–653, 2006.

[45] A. Digby, M. Towsey, B. D. Bell, and P. D. Teal. A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution*, 4(7):675–683, 2013.

[46] A. J. Doupe and P. K. Kuhl. Birdsong and human speech: common themes and mechanisms. *Annual review of neuroscience*, 22(1):567–631, 1999.

[47] T. Drugman and A. Alwan. Joint robust voicing detection and pitch estimation based on residual harmonics. In *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, August 27–31, Florence, Italy*, pages 1973–1976, 2011.

[48] S. Duan, J. Zhang, P. Roe, J. Wimmer, X. Dong, A. Truskinger, and M. Towsey. Timed probabilistic automaton: a bridge between raven and song scope for automatic species recognition. In *Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference*, pages 1519–1524. AAAI, 2013.

[49] S. Fagerlund. *Automatic recognition of bird species by their sounds.* Master's thesis, helsinki university of technology, finland, 2004.

[50] S. Fagerlund. Bird species recognition using support vector machines. *EURASIP Journal on Applied Signal Processing*, 2007(1):64–64, 2007.

[51] S. Fagerlund and U. K. Laine. New parametric representations of bird sounds for automatic classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 8247–8251, 2014.

[52] J. Feinstein, X. Yang, and S. LI. Molecular systematics and historical biogeography of the black-browed barbet species complex (megalaima oorti). *Ibis*, 150(1):40–49, 2008.

[53] G. Ferroni, R. Bonfigli, E. Principi, S. Squartini, and F. Piazza. A deep neural network approach for voice activity detection in multi-room domestic scenarios. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2015.

[54] S. T. Garnett and L. Christidis. Implications of changing species definitions for conservation purposes. *Bird Conservation International*, 17(3):187–195, 2007.

[55] D. Gelling. *Bird Song Recognition using GMMs and HMMs.* Masters dissertation university of sheffield, uk, 2010.

[56] M. Graciarena, M. Delplanche, E. Shriberg, and A. Stolcke. Bird species recognition combining acoustic and sequence modeling. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 341–344, 2011.

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

[58] A. Harma. Automatic identification of bird species based on sinusoidal modeling of syllables. In *Acoustics, Speech and Signal Processing (ICASSP), 2003 IEEE International Conference on*, volume 5, pages V–545–8 vol.5, 2003.

[59] N. Harte, S. Murphy, D. J. Kelly, and N. M. Marples. Identifying new bird species from differences in birdsong. In *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25–29, Lyon, France*, pages 2900–2904, 2013.

[60] A. J. Helbig, A. G. Knox, D. T. Parkin, G. Sangster, and M. Collinson. Guidelines for assigning species rank. *Ibis*, 144(3):518–525, 2002.

[61] J. R. Heller and J. 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.

[62] D. J. Hermes. Measurement of pitch by subharmonic summation. *The journal of the acoustical society of America*, 83(1):257–264, 1988.

[63] M. L. Isler, P. R. Isler, and B. M. Whitney. Use of vocalizations to establish species limits in antbirds (passeriformes: Thamnophilidae). *The Auk*, pages 577–590, 1998.

[64] P. Jančovič and M. Köküer. Automatic detection and recognition of tonal bird sounds in noisy environments. *EURASIP Journal on Advances in Signal Processing*, 2011(1):1–10, 2011.

[65] P. Jančovič and M. Köküer. Detection of sinusoidal signals in noise by probabilistic modelling of the spectral magnitude shape and phase continuity. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 517–520. IEEE, 2011.

[66] P. Jančovič and M. Köküer. Recognition of multiple bird species based on penalised maximum likelihood and hmm-based modelling of individual vocalisation elements. In *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association, September 8–12, San Francisco, USA*, pages 2612–2616, 2016.

[67] P. Jančovič, M. Köküer, and M. Russell. Bird species recognition from field recordings using hmm-based modelling of frequency tracks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 8252–8256, 2014.

[68] P. Jančovič, M. Köküer, M. Zakeri, and M. Russell. Unsupervised discovery of acoustic patterns in bird vocalisations employing dtw and clustering. In *21st European Signal Processing Conference (EUSIPCO 2013)*, pages 1–5. IEEE, 2013.

[69] A. Joly, H. Goëau, H. Glotin, C. Spampinato, P. Bonnet, W.-P. Vellinga, R. Planqué, A. Rauber, S. Palazzo, and B. Fisher. Lifeclef 2015: multimedia life species identification challenges. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 462–483. Springer, 2015.

[70] L. Joseph and K. E. Omland. Rowley review. phylogeography: its development and impact in australo-papuan ornithology with special reference to paraphyly in australian birds. *Emu*, 109(1):1–23, 2009.

[71] K. Kaewtip, A. Alwan, C. O'Reilly, and C. E. Taylor. A robust automatic birdsong phrase classification: A template-based approach. *The Journal of the Acoustical Society of America*, 140(5):3691–3701, 2016.

[72] K. Kaewtip, L. N. Tan, A. Alwan, and C. E. Taylor. A robust automatic bird phrase classifier using dynamic time-warping with prominent region identification. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 768–772, 2013.

[73] H. Kawahara, J. Estill, and O. Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *MAVEBA*, pages 59–64, 2001.

[74] H. Kawahara, H. Katayose, A. De Cheveigné, and R. D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. In *EuroSpeech*, volume 99, pages 2781–2784, 1999.

[75] F. Kelly. *Automatic Recognition of Ageing Speakers*. Ph.d. dissertation, trinity college dublin, ireland, 2014.

[76] R. D. Kent. *The MIT encyclopedia of communication disorders*. MIT Press, 2004.

[77] A. Kershenbaum, D. T. Blumstein, M. A. Roch, Ç. Akçay, G. Backus, M. A. Bee, K. Bohn, Y. Cao, G. Carter, and C. Cäsar. Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews*, 2014.

[78] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40, 2010.

[79] J. A. Kogan and D. 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.

[80] D. Kohlsdorf, C. Mason, D. Herzing, and T. Starner. Probabilistic extraction and discovery of fundamental units in dolphin whistles. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 8242–8246, 2014.

[81] R. Kojima, O. Sugiyama, R. Suzuki, K. Nakadai, and C. E. Taylor. Semi-automatic bird song analysis by spatial-cue-based integration of sound source detection, localization, separation, and identification. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 1287–1292. IEEE, 2016.

[82] A. H. Krakauer, M. Tyrrell, K. Lehmann, N. Losin, F. Goller, and G. L. Patricelli. Vocal and anatomical evidence for two-voiced sound production in the greater sage-grouse centrocercus urophasianus. *Journal of Experimental Biology*, 212(22):3719–3727, 2009.

[83] S. Kullback. *Information theory and statistics*. Courier Corporation, Dover Publications Inc. New York, 1968 republished 1997.

[84] R. F. Lachlan. Luscinia, 2016. http://luscinia.sourceforge.net/ (accessed 29 Sept 2016).

[85] R. F. Lachlan, M. N. Verzijden, C. S. Bernard, P.-P. Jonker, B. Koese, S. Jaarsma, W. Spoor, P. J. Slater, and C. Ten Cate. The progressive loss of syntactical structure in bird song along an island colonization chain. *Current Biology*, 23(19):1896–1901, 2013.

[86] F. Lambert and P. Rasmussen. A new scops owl from sangihe island, indonesia. *BULLETIN-BRITISH ORNITHOLOGISTS CLUB*, 118:204–216, 1998.

[87] B. S. Lee and D. P. W. Ellis. Noise robust pitch tracking by subband autocorrelation classification. In *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association, September 9–13, Portland, Oregan, USA*, pages 707–710, 2012.

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

[89] C.-H. Lee, S.-B. Hsu, J.-L. Shih, and C.-H. Chou. Continuous birdsong recognition using gaussian mixture modeling of image shape features. *Multimedia, IEEE Transactions on*, 15(2):454–464, 2013.

[90] B. Library. British national sound archive, 2016. http://explore.bl.uk/primo_library/libweb/ (accessed 26 Nov 2016).

[91] M. Lindermuth. Harma syllable segmentation, 2010. http://uk.mathworks.com/matlabcentral/fileexchange/29261-harma-syllable-segmentation (accessed 03 Nov 2016).

[92] I. Luengo, I. Saratxaga, E. Navas, I. Hernáez, J. Sanchez, and I. Sainz. Evaluation of pitch detection algorithms under real conditions. In *Acoustics, Speech and Signal Processing (ICASSP), 2007 IEEE International Conference on*, volume 4, pages IV–1057–IV–1060, 2007.

[93] G. M. Mace. The role of taxonomy in species conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1444):711–719, 2004.

[94] J. Mallet. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1506):2971–2986, 2008.

[95] Y. Mandelblat-Cerf and M. S. Fee. An automated procedure for evaluating song imitation. *PloS one*, 9(5):e96484, 2014.

[96] P. R. Marler and H. Slabbekoorn. *Nature's music: the science of birdsong*. Academic Press, 2004.

[97] T. A. Marques, L. Thomas, S. W. Martin, D. K. Mellinger, J. A. Ward, D. J. Moretti, D. Harris, and P. L. Tyack. Estimating animal population density using passive acoustics. *Biological Reviews*, 88(2):287–309, 2013.

[98] J. Martens, S. Eck, M. Packert, and Y. Sun. Methods of systematic and taxonomic research on passerine birds: the timely example of the seicercus burkii complex (sylviidae). *Bonner zoologische Beitrage*, 51(2/3):109–118, 2003.

[99] R. M. May. Taxonomy as destiny. *Nature*, 347(6289):129–130, 1990.

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

[101] B. D. McKay, M. B. J. Reynolds, W. K. Hayes, and D. S. Lee. Evidence for the species status of the bahama yellow-throated warbler (dendroica "dominica" flavescens). *The Auk*, 127(4):932–939, 2010.

[102] M. Mehrabani, H. Boril, and J. H. Hansen. Dialect distance assessment method based on comparison of pitch pattern statistical models. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5158–5161, 2010.

[103] M. Mehrabani and J. H. Hansen. Automatic analysis of dialect/language sets. *International Journal of Speech Technology*, pages 1–10, 2015.

[104] S. Meignier and T. Merlin. Lium spkdiarization: an open source toolkit for diarization. In *Carnegie Mellon University (CMU) Sphinx Users and Developers Workshop (SPUD) Workshop*, volume 2010, 2010.

[105] C. D. Meliza, S. C. Keen, and D. R. Rubenstein. Pitch-and spectral-based dynamic time warping methods for comparing field recordings of harmonic avian vocalizations. *The Journal of the Acoustical Society of America*, 134(2):1407–1415, 2013.

[106] D. B. Miller. Two-voice phenomenon in birds: further evidence. *The Auk*, pages 567–572, 1977.

[107] I. R. Murray and J. L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108, 1993.

[108] S. Nakagawa and I. C. Cuthill. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(4):591–605, 2007.

[109] L. Neal, F. Briggs, R. Raich, and X. Z. Fern. Time-frequency segmentation of bird song in noisy acoustic environments. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2012–2015. IEEE, 2011.

[110] A. M. Noll. Cepstrum pitch determination. *The journal of the acoustical society of America*, 41(2):293–309, 1967.

[111] S. Nowicki and R. R. Capranica. Bilateral syringeal coupling during phonation of a songbird. *The Journal of neuroscience*, 6(12):3595–3610, 1986.

[112] S. Nowicki and P. Marler. How do birds sing? *Music Perception: An Interdisciplinary Journal*, 5(4):391–426, 1988.

[113] H. G. Okuno and K. Nakadai. Robot audition: Its rise and perspectives. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5610–5614. IEEE, 2015.

[114] C. O'Reilly. Bird synthesis database, 2016. http://www.mee.tcd.ie/~sigmedia/REsources/SynthBirdsData (accessed 29 Sept 2016).

[115] C. O'Reilly. Bird synthesis listening survey, 2016. http://www.surveygizmo.com/s3/2570269/Bird-synth-comparison (accessed 29 Sept 2016).

[116] C. O'Reilly, N. M. Marples, D. J. Kelly, and N. Harte. Quantifying difference in vocalizations of bird populations. In *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6–10, Dresden, Germany*, pages 3417–3421, 2015.

[117] C. O'Reilly, N. M. Marples, D. J. Kelly, and N. Harte. Yin-bird: Improved pitch tracking for bird vocalisations. In *INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association, September 8–12, San Francisco, USA*, pages 2641–2645, 2016.

[118] C. L. o. Ornithology. Macaulay library, 2016. http://macaulaylibrary.org/ (accessed 26 Nov 2016).

[119] M. Päckert, J. Martens, J. Kosuch, A. A. Nazarenko, and M. Veith. Phylogenetic signal in the song of crests and kinglets (aves: Regulus). *Evolution*, 57(3):616–629, 2003.

[120] M. A. Patten, P. Unitt, and F. Sheldon. Diagnosability versus mean differences of sage sparrow subspecies. *The Auk*, 119(1):26–35, 2002.

[121] N. Pieplow and A. Spencer. Earbirding.com - the seven basic tone qualities and how to read spectrograms?, 2013. http://earbirding.com/blog/archives/4621 http://earbirding.com/blog/specs (accessed 22 Nov 2016).

[122] R. Pieraccini. From audrey to siri. is speech recognition a solved problem?, 2012. http://www.icsi.berkeley.edu/pubs/speech/audreytosiri12.pdf(accessed 18 Jan 2017).

[123] I. Potamitis, S. Ntalampiras, O. Jahn, and K. Riede. Automatic bird sound detection in long real-field recordings: Applications and tools. *Applied Acoustics*, 80(0):1–9, 2014.

[124] A. Purvis, J. L. Gittleman, and T. Brooks. *Phylogeny and conservation - Species: demarcation and diversity, 57-75*. Cambridge University Press, 2005.

[125] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE Acoustics, Speech, and Signal Processing Magazine*, 3(1):4–16, 1986.

[126] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[127] J. Ramirez, J. M. Górriz, and J. C. Segura. *Voice activity detection. fundamentals and speech recognition system robustness*. INTECH Open Access Publisher, 2007.

[128] L. Ranjard and H. A. Ross. Unsupervised bird song syllable classification using evolving neural networks. *The Journal of the Acoustical Society of America*, 123(6):4358–4368, 2008.

[129] P. Rasmussen, J. Wardill, F. Lambert, and J. Riley. On the specific status of the sangihe white-eye zosterops nehrkorni, and the taxonomy of the black-crowned white-eye z. atrifrons complex. *Forktail*, pages 69–82, 2000.

[130] J. Remsen Jr. Pattern, process, and rigor meet classification. *The Auk*, 122(2):403–413, 2005.

[131] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

[132] F. E. Rheindt, J. A. Norman, and L. Christidis. Dna evidence shows vocalizations to be a better indicator of taxonomic limits than plumage patterns in zimmerius tyrant-flycatchers. *Molecular Phylogenetics and Evolution*, 48(1):150–156, 2008.

[133] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier. An open-source state-of-the-art toolbox for broadcast news diarization. In *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25–29, Lyon, France*, 2013.

[134] O. A. Ryder. Species conservation and systematics: the dilemma of subspecies. *Trends in Ecology & Evolution*, 1(1):9–10, 1986.

[135] J. Saeedi, S. M. Ahadi, and K. Faez. Robust voice activity detection directed by noise classification. *Signal, Image and Video Processing*, 9(3):561–572, 2015. https://uk.mathworks.com/matlabcentral/fileexchange/39343-voice-activity-detection-directed-by-noise-classification/content/vad_directed_by_noise_classification.m (accessed 18 Jan 2017).

[136] M. Sakamoto and T. Saito. Speaker recognizability evaluation of a voicefont-based text-to-speech system. In *INTERSPEECH 2002 – 3rd Annual Conference of the International Speech Communication Association, September 16–20, Denver, Colorado, USA*, 2002.

[137] G. Sangster, B. F. King, P. Verbelen, and C. R. Trainor. A new owl species of the genus otus (aves: Strigidae) from lombok, indonesia. *PloS one*, 8(2):e53712, 2013.

[138] K. Sasahara, M. L. Cody, D. Cohen, and C. E. Taylor. Structural design principles of complex bird songs: a network-based approach. *PLoS ONE*, 7:e44436, 2012.

[139] P. Scheifele, S. Andrew, R. Cooper, M. Darre, F. Musiek, and L. Max. Indication of a lombard vocal response in the st. lawrence river beluga. *The Journal of the Acoustical Society of America*, 117(3):1486–1492, 2005.

[140] M. R. Schroeder. Period histogram and product spectrum: New methods for fundamental-frequency measurement. *The Journal of the Acoustical Society of America*, 43(4):829–834, 1968.

[141] Sennheiser. Sennheiser me62 omni-directional condenser microphone capsule with k6 power supply, 2013. http://www.bhphotovideo.com/c/product/73085-REG/Sennheiser_ME62_ME62_Omni_Mic.html (accessed on 22nd April 2014).

[142] X. Serra. *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*. Ph.d. dissertation, stanford university, usa, 1989.

[143] X. Serra. Sinusoid plus residual python code, 2007. https://github.com/MTG/sms-tools/blob/master/software/models_interface/sprModel_function.py (accessed 03 November 2016).

[144] W. A. Sethares. *Tuning, timbre, spectrum, scale*. Springer Science & Business Media, 2005.

[145] S. Shaw. *An Evaluation of Birdsong Recognition Techniques*. Undergraduate dissertation, university of sheffeild , uk, 2011.

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

[147] E. Sprengel, Y. Martin Jaggi, and T. Hofmann. Audio based bird species identification using deep learning techniques. *Working notes of CLEF*, 2016.

[148] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin. Bird detection in audio: a survey and a challenge. In *IEEE International Workshop on Machine Learning for Signal Processing, SEPT. 13–16, 2016, Salerno, Italy*, 2016.

[149] C. B. Sturdy and R. Mooney. Bird communication: Two voices are better than one. *Current Biology*, 10(17):R634–R636, 2000.

[150] J. Sueur, A. Farina, A. Gasc, N. Pieretti, and S. Pavoine. Acoustic indices for biodiversity assessment and landscape investigation. *Acta Acustica united with Acustica*, 100(4):772–781, 2014.

[151] X. Sun. Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages I–333–I–336. IEEE, 2002.

[152] D. Talkin. A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis Journal*, 495:518, 1995.

[153] L. N. Tan, A. Alwan, G. Kossan, M. L. Cody, and C. E. Taylor. Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training dataa). *The Journal of the Acoustical Society of America*, 137(3):1069–1080, 2015.

[154] L. N. Tan, K. Kaewtip, M. L. Cody, C. E. Taylor, and A. Alwan. Evaluation of a sparse representation-based classifier for bird phrase classification under limited data conditions. In *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association, September 9–13, Portland, Oregan, USA*, pages 2522–2525, 2012.

[155] L. N. Tan, G. Kossan, M. L. Cody, C. E. Taylor, and A. Alwan. A sparse representation-based classifier for in-set bird phrase verification and classification with limited training data. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 763–767, 2013.

[156] O. Tchernichovski, E. Kashtelyan, D. Swigger, and P. P. Mitra. Sound analysis pro (sap) software download, 2011. http://soundanalysispro.com/ (accessed 22 Nov 2016).

[157] O. Tchernichovski, F. Nottebohm, C. E. Ho, B. Pesaran, and P. P. Mitra. A procedure for an automated measurement of song similarity. *Animal Behaviour*, 59(6):1167–1176, 2000.

[158] Telinga. Telinga 1 mm flexible dish, 2013. http://www.telinga.com/products/dishes/ (accessed 22 Apr 2014).

[159] N. S. Thompson, K. LeDoux, and K. Moody. A system for describing bird song units. *Bioacoustics: The International Journal of Animal Sound and its Recording*, 5(4):267–279, 1994.

[160] J. A. Tobias, N. Seddon, C. N. Spottiswoode, J. D. Pilgrim, L. D. Fishpool, and N. J. Collar. Quantitative criteria for species delimitation. *The International Journal of Avian Science (IBIS)*, 152(4):724–746, 2010.

[161] J. A. Tobias, N. Seddon, C. N. Spottiswoode, J. D. Pilgrim, L. D. Fishpool, and N. J. Collar. Quantitative criteria for species delimitation - supporting information link, 2010. http://onlinelibrary.wiley.com/wol1/doi/10.1111/j.1474-919X.2010.01051.x/suppinfo (accessed 29 Nov 2016).

[162] B. P. Tóth and B. Czeba. Convolutional neural networks for large-scale bird song classification in noisy environment. In *Working notes of CLEF conference*, 2016.

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

[164] S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, 2006.

[165] V. M. Trifa, A. N. Kirschel, C. E. Taylor, and E. E. Vallejo. Automated species recognition of antbirds in a mexican rainforest using hidden markov models. *The Journal of the Acoustical Society of America*, 123(4):2424–2431, 2008.

[166] B. P. Vellinga and Willem-Pier. Xeno-canto.org online database, accessed on 07 august, 2014. http://www.xeno-canto.org (accessed 22 Nov 2016).

[167] E. Vilches, I. A. Escobar, E. E. Vallejo, and C. E. Taylor. Data mining applied to acoustic bird species recognition. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 400–403, 2006.

[168] C. wa Maina. Audio diarization for biodiversity monitoring. In *AFRICON, 2015*, pages 1–5. IEEE, 2015.

[169] A. Wang. An industrial strength audio search algorithm. In *ISMIR*, pages 7–13. Washington, DC, 2003.

[170] M. Wang and M. Lin. An analysis of pitch in chinese spontaneous speech. In *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, 2004.

[171] C. Wei and A. Alwan. Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend. In *Acoustics, Speech and Signal Processing (ICASSP), 2009 IEEE International Conference on*, pages 3969–3972, 2009.

[172] C. Wei and A. Alwan. Fbem: A filter bank em algorithm for the joint optimization of features and acoustic model parameters in bird call classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 1993–1996, 2012.

[173] C. M. U. Wildlife Acoustics Inc. Song scope software, 2011. http://www.wildlifeacoustics.com/ (accessed 07 Dec 2016).

[174] W. Woodland Park Zoo, Seattle. Bird basics presentation, 2013. http://slideplayer.com/slide/5722819/ (accessed 16 Dec 2016).

[175] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge University Publishing Department, 2002. http://htk.eng.cam.ac.uk (accessed 17 Jan 2017).

[176] X.-L. Zhang and D. Wang. Boosting contextual information for deep neural network based voice activity detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):252–264, 2016.

[177] S. A. Zollinger, T. Riede, and R. A. Suthers. Two-voice complexity from a single side of the syrinx in northern mockingbird mimus polyglottos vocalizations. *Journal of Experimental Biology*, 211(12):1978–1991, 2008.

[178] ZOOM. Zoom handy recorder h2, 2013. http://www.samsontech.com/site_media/legacy_docs/H2_user_manual.pdf (accessed on 22nd Apr 2014).