AUTOMATIC BIRD SPECIES IDENTIFICATION EMPLOYING AN UNSUPERVISED DISCOVERY OF VOCALISATION UNITS

By

Masoud Zakeri

A Thesis Submitted to
The University of Birmingham
for the Degree of

Doctor of Philosophy

School of Electronic, Electrical
and System Engineering
College of Engineering
and Physical Sciences
The University of Birmingham
July 2017

Abstract

An automatic analysis of bird vocalisations for the identification of bird species, the study of their behaviour and their means of communication is important for a better understanding of the environment in which we are living and in the context of environmental protection. Currently, the highly skilled bird surveyors who observe and monitor bird species identify them visually and more often acoustically, especially in habitats with low visibility. The high variability of vocalisations within different individuals makes species' identification challenging for bird surveyors. In addition, there are many amateur bird watchers, all of whom use book-based bird guides in order to identify birds. A device that could be left on site and could screen bird songs and identify the bird species, would provide a more complete survey than is possible with a small number of surveyor visits, and furthermore would reduce the cost. Hence, the availability of a reliable automatic bird identification system through their vocalisations, would be of great interest to professionals and amateurs alike.

A part of this thesis provides a biological survey on the scientific theories of the study of bird vocalisation and corresponding singing behaviours.

Another section of this thesis aims to discover a set of element patterns produced by each bird species in a large corpus of the natural field recordings. This is performed in three steps. First, a set of frequency tracks segments are extracted from each vocalisation acoustic signal, by using a sinusoidal detection method. Then in the next step, a novel approach is proposed to search for partial and multiple matchings between each pair of detected segments, using a modified DTW. This is implemented by several parallel DTW searches; each performs from a different starting point on one of the segments and can start anywhere on the other segment. Then in the last step, these DTW searching outputs consisting of the obtained pairwise partial similarity score with the corresponding matching path of the entire detected segments, are used in a novel presented hierarchical clustering approach to group all the homogeneous structured segments into a set of distinct element-based vocalisation clusters. The obtained result has demonstrated the good coherence of element patterns within each cluster and clearly distinctive patterns among the clusters.

This thesis aims to develop an automatic system for the identification of bird species from natural field recordings. Two HMM based recognition systems on the frequency tracks segments, which were detected, are presented in this research. In the baseline system the entire segments of each bird species are modelled with a single HMM model. Then in a novel bird identification approach, a element-based HMM recognition system is presented and each individual vocalisation element is modelled with a single HMM. To build the corresponding element HMMs, as there is no such labelling information available for the vocalisation elements in the real natural recorded corpus, the output of the above-mentioned unsupervised discovery method is used as the training label information. Experiments have been performed on over 38 hours of natural field recordings, consisting of 48 bird species. Evaluations have been demonstrated where the proposed element-based HMM system obtained a recognition accuracy of over 93% by using 3 seconds of detected signal and over 39% recognition error rate reduction, compared to the baseline HMM system of the same complexity.

This thesis is dedicated to my parents

for their endless love, support

and encouragement

Acknowledgements

First of all, I would like to thank my supervisor Dr Peter Jančovič for all the help, support and encouragement that he gave me throughout my studies and it was a pleasure to work with him over the past years.

This project was carried out on a data set provided by Borror Laboratory of Bioacoustics and without their generosity and sharing their data set this project would have been much harder to achieve.

All the staff and colleagues in the DSVP research group made me feel welcome over the past few years while I was studying and always helped me whenever I needed; I will miss them all, especially my colleagues Abualsoud, Saeid and Phil.

The members of staff in the School of Engineering were always very supportive and encouraging. The list is enormous but in particular, I would like to thank Prof. Peter Gardner, Prof. Martin Russell, Dr Münevver Köküer, Prof. Chris Baber, Mary Winkles, Samantha McCauley, Phil Atkins, Ben Clarke, Andy Dunn and other members of staff in the School.

My deepest gratitude goes to my family; especially my dear siblings Nasrin, Mohsen and Zohreh, who are very special to me and have always been by my side. My father and mother always helped me throughout my life and without their support I would never have been able to reach this stage in my life. Their unconditional support and love always lights my way to success.

Finally, I would like to acknowledge the help and support of my friends Farzad, Hesam, Mohammadreza, Mohammadhossein Zoualfaghari, Mohammadreza Zolfaghari, Laleh Akbarzadeh, Mohammadhossein Moghimi and anyone else who helped me during my studies over the past few years.

This thesis was copy edited for conventions of language, spelling and grammar by Janet's Proofreading Service.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Introduction

Identification of bird species with an automatic analysis of bird vocalisations has many potential applications in ecology, nature conservation monitoring, and vocal behavioural studies. Scientific research, such as [1], on different bird species demonstrates that birds are a good sign of the state of our living environment; since they are widely distributed, they react rapidly to variations in environmental conditions such as climate change. There are around 700 environmental consultancy companies in the UK which cumulatively carry out thousands of bird survey days each year. These bird surveys are performed as a standard part of the environmental impact assessment of sites due for development, or in the context of conservation. Currently, these surveys are traditionally performed by a large number of highly skilled bird surveyors and ornithologists, who observe and monitor bird species, identify them visually and more often acoustically, especially in habitats with low visibility. The wide variability of vocalisations within different individuals makes species' identification challenging for bird surveyors. Also, surveys performed by different people may not be entirely comparable, as each surveyor may have a different level of skills. Hence, this is an area where automated identification of bird species from their vocalisation systems could offer huge advantages. Use of an automated analysis of bird vocalisations in acoustical monitoring systems, provides more efficient observation with lower surveying costs. Moreover, there are many amateur birdwatchers, all of whom use textual bird guides, such as books and magazines, in order to identify birds. Thus, an automatic bird recognition (software) application would also provide an educational and entertaining alternative for both amateur and professional birdwatchers.

Automatic processing of bird vocalisations is a relatively recent research field [2, 3, 4] and many previous related approaches are developed based on the techniques which are used in speech and language signal processing. The main goal of this thesis is to present an automatic system for the identification of bird species from natural field recordings.

1.2 Major Contributions

The research introduced in this thesis provides original contributions to the field of automatic processing and classification of bird acoustic signals. The major contributions of this thesis are summarised as follows:

1- The manual annotation of a large corpus of bird vocalisation recording files, which are recorded in natural field environments. In order to evaluate the performance of the automatic segmentation and feature extraction procedures, and also to allow other researchers to perform and evaluate comparative experiments with the real field corpus, the whole data is inspected manually to obtain the acoustic event-based label file for each vocalisation recording. For the sake of refining or modifying the acoustic events in the provided labelling files, a user interface (a MATLAB script) has been supplied among the annotation files (Chapter 4).

2- Development of a partial DTW similarity calculation algorithm to search for the partial and multiple matchings between a given pair of (temporal) sequences. In order to find the partial paths, the proposed algorithm employs a variant of DTW in several searching procedures; where each search considers a different time-stamp on one of the sequences and allows the DTW alignment path to start and end anywhere on the second sequence. The obtained pairwise partial similarity path can be represented by the partial similarity score and the corresponding time-stamps of the detected partial path on both sequences. This novel method is explained in Chapter 6.

3- Development of a novel hierarchical clustering algorithm that employs the (obtained) partial similarity information, including the similarity scores and the corresponding time-stamps of the partial path, of the entire vocalisation segments of each bird species in order to group together all structurally similar segments. Several rules and conditions are used in this method to control the merging decisions. In other words, the merge decisions of the clustering procedure are always based on further investigations of the prospective likeness structure of the group, including both the merging objects. This clustering algorithm, along with the partial DTW similarity calculation algorithm are used in an unsupervised manner to discover a set of distinct vocalisation elements for each bird species in the data set. This novel algorithm is explained in Chapter 6.

4- Development of a novel automatic bird species identification system based on HMM modelling of individual element vocalisation units. In this approach, instead of employing a single HMM model for each bird species, a single HMM is used to model each type of vocalisation pattern that is available in each particular bird species. As there is no further element-level information available among the natural field recordings, training the element-based models is not practical. Hence, this proposed system employs the outcome of hierarchical clustering algorithm, as label information to train the HMM models. The experimental evaluations on the proposed identification approach demonstrate that the recognition accuracy is significantly improved with the error rate reduction between 39% and 48% in comparison with the proposed baseline system. This novel recognition system is explained in Chapter 7.

1.3 Thesis structure

This thesis is organised in eight chapters with two appendices, A and B. Figure 1.1 shows a block diagram of the thesis structure. The following sections summarise the information provided in each chapter.

1.3.1 Chapter 2 - Literature review of bird sound recognition systems and techniques

This chapter describes the background techniques of audio pattern processing for developing a typical automatic audio recognition system.

1.3.2 Chapter 3 - The study of bird vocalisation

This chapter introduces the basic biological theories of bird vocalisations by studying the communication and singing behaviours of typical passerine birds; this is followed by a description of the corresponding song terminology and the scientific theories about vocal learning and development procedures in young birds.

1.3.3 Chapter 4 - Bird vocalisation corpus

The first part of this chapter introduces a brief literature review of the large, available bird vocalisation archives, followed by a description of the data set which has been used in recent international bird classification challenges. In the second part of the chapter, firstly the database which is used in all the experimental evaluations of this study is presented; then each vocalisation audio file of this large database, which was recorded in the natural habitats of the birds, is manually annotated (classified) into the 12 pre-defined sound events. A part of this annotated data was used in Chapter 5 to evaluate the performance of the entire bird detection system in terms of segmentation and feature extraction tasks.

1.3.4 Chapter 5 - Segmentation and estimation of acoustic features for bird vocalisation

In this chapter, the proposed sinusoidal detection approach in [5, 6] has been employed with some modification, as manner of automatic segmentation and a feature extraction step, to decompose the entire acoustic recordings of the data into the set of distinct frequency components (frequency track segments) to characterise bird tonal vocalisation. Then the detected segments are used as temporal sequences for the further processing stages in the following Chapters 6 and 7.

1.3.5 Chapter 6 - Unsupervised discovery of acoustic elements in bird vocalisation

This chapter presents an approach for unsupervised discovery of acoustic elements in bird vocalisations recorded in real world natural environments. This proposed system is comprised of two steps, obtaining the pairwise partial similarity paths and clustering the entire segments by using the obtained partial matching information. The proposed approach employs the detected frequency track segments to obtain a set of individual element vocalisation patterns for each bird species. The obtained result then is used in the next chapter as element-level labelling information.

1.3.6 Chapter 7 - An automatic HMM-based bird sound recognition system

This chapter presents two automatic HMM based approaches as baseline and element-based systems, for identification of bird species from the natural field recordings, by using the detected frequency tracks as temporal sequences. The obtained element-level labelling information in Chapter 6 is used to train the HMMs in the novel proposed element-based recognition system. Finally, experimental evaluations on both systems are provided, a long with presenting a review of previous studies on bird vocalisation recognition systems.

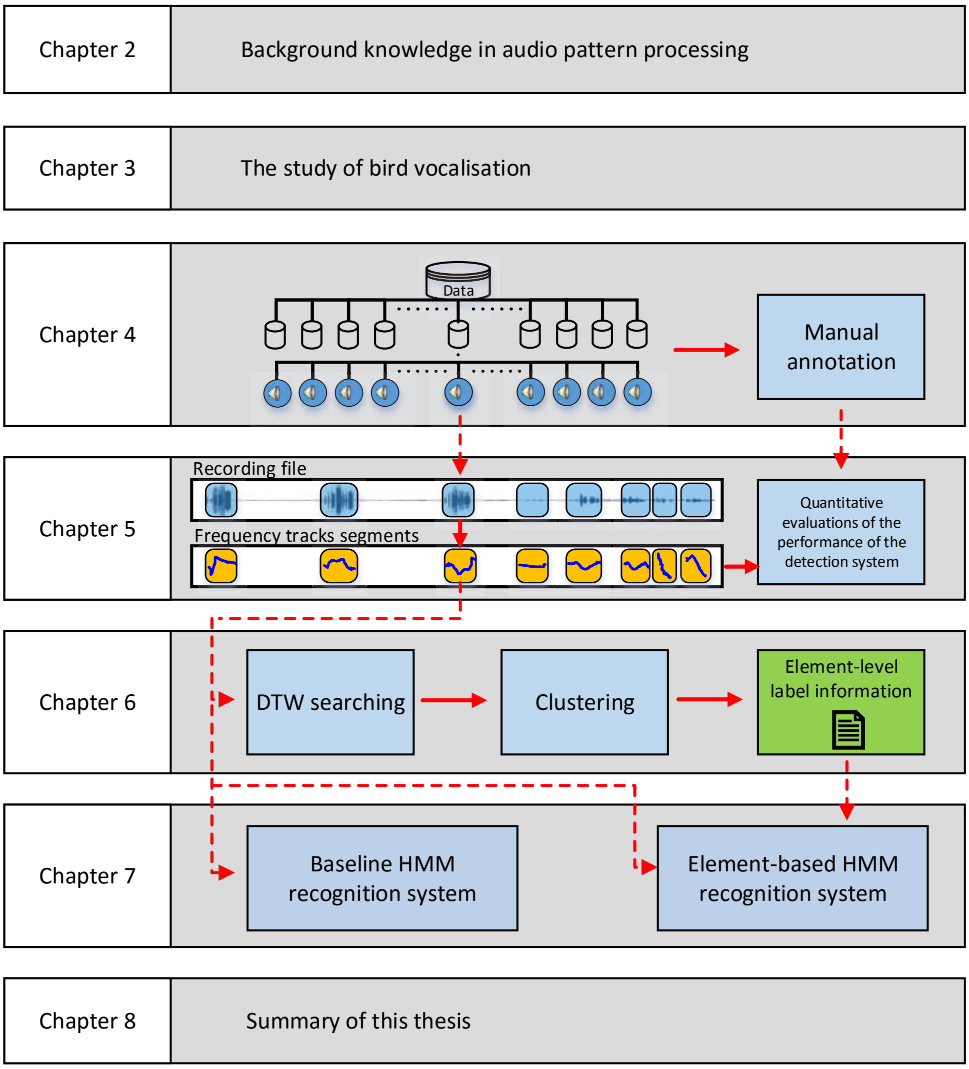

Chapter 2
Background knowledge in audio pattern processing
Chapter 3
The study of bird vocalisation
Chapter 4
Data
Manual annotation
Chapter 5
Recording file
Frequency tracks segments
Quantitative evaluations of the performance of the detection system
Chapter 6
DTW searching
Clustering
Element-level label information
Chapter 7
Baseline HMM recognition system
Element-based HMM recognition system
Chapter 8
Summary of this thesis

Figure 1.1: Block diagram of the thesis structure

1.3.7 Chapter 8 - Conclusion

The final chapter summarises the contributions and draws a conclusion of the thesis in addition to anticipating the possible future directions of the work.

1.4 List of publications

The list of publications of the author is as follows:

- Jancovic P, Kokuer M, Zakeri M, Russell M. Unsupervised discovery of acoustic patterns in bird vocalisations employing DTW and clustering. In: Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European. IEEE; 2013. p. 1-5.
- Jancovic P, Zakeri M, Kokuer M, Russell M. HMM-based modelling of individual syllables for bird species recognition from audio field recordings. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on IEEE; 2015. p. 768-772.
- Jancovic P, Kokuer M, Zakeri M, Russell M. Improving acoustic and incorporating duration modelling into HMM-based recognition of bird species from audio field recordings. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on IEEE; 2016. p. 559-563.

1.5 Summary

This chapter presents a brief introduction to this research, its aims and objectives and its contributions, including a brief summary of each chapter of the thesis; followed by the list of publications of the author.

Chapter 2

LITERATURE REVIEW ON BIRD SOUND RECOGNITION SYSTEMS AND TECHNIQUES

2.1 Introduction

The field of automatic bird sound recognition processing is a particular application area that has developed from the more general fields of audio and speech processing and pattern recognition. The literature in this chapter is described in two parts. In the first part, literature on common audio pattern processing methods for developing an example of automatic audio recognition system is presented. In the second part, a brief review of the literature on hierarchical clustering methods is presented.

2.2 An overview of an example of the automatic audio recognition system

This section provides a brief summary of a given example of the automatic audio recognition system. First, all the available data is split into two parts: training and testing data. As seen in Figure 2.1, the recognition system example is performed in two main phases, training and testing, with the following main components in each stage:

Segmentation: the goal in this phase is to remove the parts, which are silence or void of audio, from the continuous input audio signal. The consequence of this is that the signal is divided into isolated temporal segments. Hence, the segmentation also performs data reduction by omitting periods of time in continuous recordings when audio events are not present.

Feature extraction: in this stage, each detected audio segment will be transformed into a sequence of features (parametric representations). The goal in feature extraction is to extract relevant information from the input signal into a compact set of features that is then used to distinguish between different classes.

Modelling: in the training phase, once individually distinct features have been extracted from a signal, a model is used to represent each sound class based on a set of features extracted from the training set of data.

Recognition: in the testing phase, a classifier is used to compare the features extracted from an unknown signal with the trained models of each class, in such a way as to find the most likely model which corresponds to the recognition result.

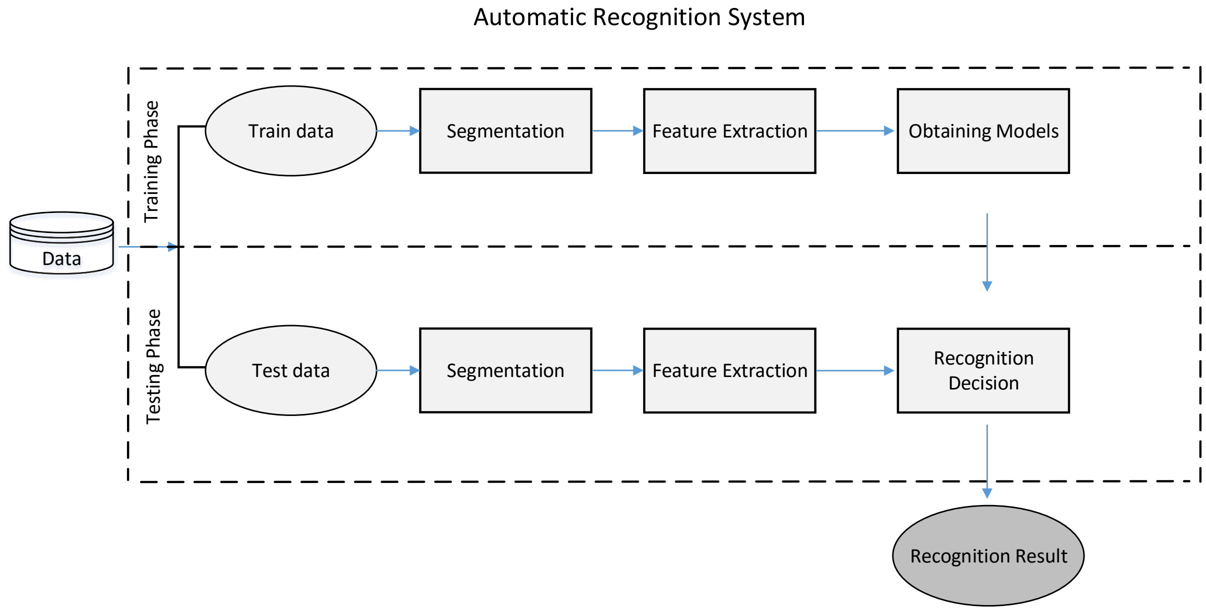


Figure 2.1: A block diagram of an example of an automatic audio recognition system

Each of the above components of the recognition system forms its own distinct area of research and faces its own set of basic challenges. Hence, the following sections present a review of the common approaches for signal segmentation, feature extraction and acoustic modelling (classification) in the literature.

2.3 Signal segmentation

Signal segmentation is performed after the data preparation stage and aims to split the continuous acoustic signal into smaller isolated units where sound events are present. Accurate segmentation is an essential step in the recognition system, in particular for bird vocalisations recorded in their natural environments containing other background sounds; inaccurate segmentation can increase the occurrence of misclassifications. This procedure can be done manually or automatically. Although manual segmentation was used in early research on bird sound analysis [7, 8, 4, 9], it is only feasible when using a small set of data. Therefore, it is essential to use an automatic detection method to segment all of the large sets of data.

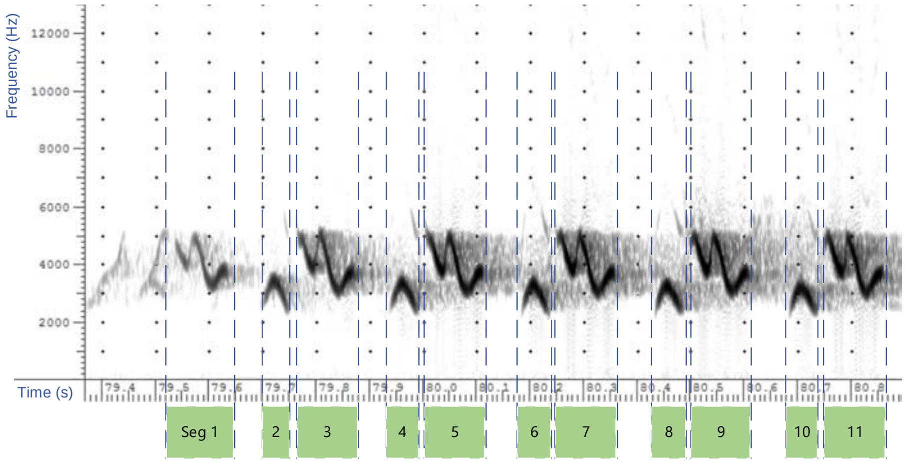


Figure 2.2: Example of splitting a continuous bird vocalisation signal into smaller segments (elements).

Most works in audio recognition processing (including bird sounds) have performed automatic segmentation by using an energy-based detector [10, 11, 2, 12], i.e. computing an energy envelope for a sound signal, to keep only the parts where the energy is above a set threshold and remove the parts which have low energy. Furthermore, in most recent works on bird sound analysis [2, 9, 13, 11], the isolated segments obtained from the segmentation process can be seen to correspond to individual elements of bird vocalisation (see Figure 2.2). In Chapter 3, more about these elements as the building blocks of bird vocalisation will be discussed.

2.4 Feature extraction

The aim of feature extraction is to compress the audio signal into a distinct form that characterises the important information of each sound event. As the result of the subsequent classification procedure relies on the feature extraction section, the features should not be sensitive to some influences such as noise and should be able to discriminate between various acoustic events.

The most popular acoustic features that are used in bioacoustics signal processing approaches, including bird vocalisation approaches, are inspired merely from the feature extraction techniques which are developed in speech and spoken language processing [14]. Mel-Frequency Cepstral Coefficients (MFCCs) [15, 16] and Linear Prediction Cepstral Coefficients (LPCCs) [17] are two of the most common frame-based features where both methods represent the spectral envelope of the signal.

MFCCs represent the discrete cosine transform (DCT) of the log-spectral power of the signal mapped onto the non-linear Mel frequency scale, where the relation between Mel and frequency f in kHz is derived as;

$$Mel = 1000log_2(1+f) \tag{2.1}$$

Figure 2.3 summarizes the procedure of MFCCs' and LPCCs' extraction.

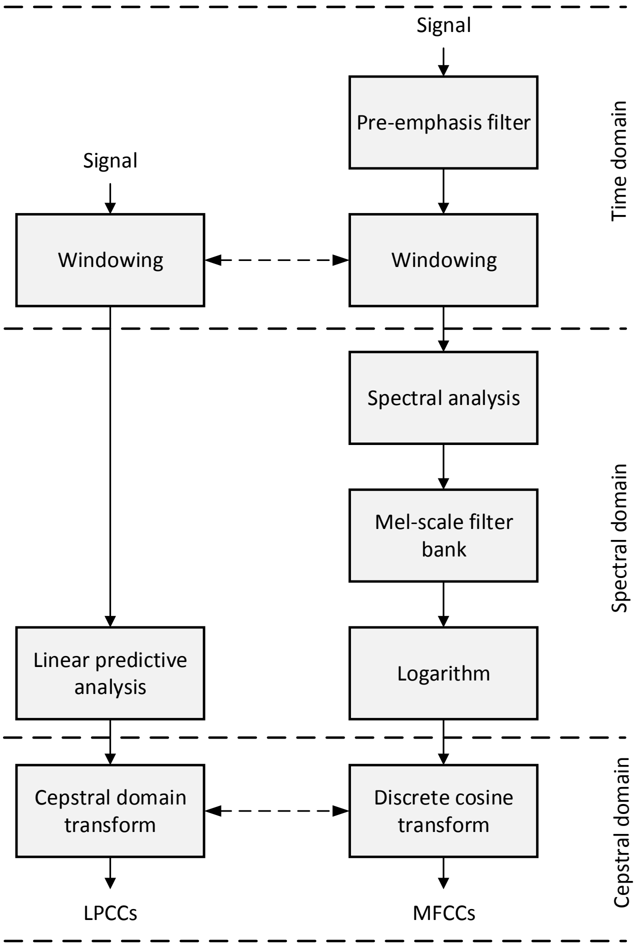


Figure 2.3: Feature extraction process for LPCCs and MFCCs. Dashed lines indicate equivalents processes.

2.5 Acoustic modelling

After extracting individual distinct features from a signal, a classifier is used to characterize the feature sets and acquire a model for each individual (training phase). It is then used in the testing stage to compare a given series of features with the saved reference templates to settle on a choice of identity [18]. In general, each classifier is composed of one or several classification methods where the main goal is to distinguish and classify a series of unknown sequence objects, based on the classification measurements [19].

Classification tasks in audio and speech pattern recognition fields can comprise of either identification or verification. In identification, an input signal is looked at against a library of template signals from known acoustic objects (classes), where the best match is chosen as the identity of the test object [20]. Verification is used exclusively to verify the asserted identity of a given acoustic signal, by comparing the corresponding feature vectors with the stored trained templates to accept or reject the claimed identity [20]. In birds' recognition exercises both of these classification tasks are performed and in this research bird species' identification is considered as a means of classification.

There is a wide range of classification methods available in the context of machine learning and pattern recognition. Some classifiers, such as the hidden Markov model (HMM) and Gaussian mixture models (GMM), are the generative approaches on which the data distribution is modelled in order to categorise the given signal; whereas others, such as support vector machines (SVM) and artificial neural networks, are the discriminative approaches where the classification decision relies on the characteristics of the data. In the following sections, the fundamentals of two different widely used classifiers in audio and speech signal processing, GMMs and HMMs, are outlined.

2.5.1 Gaussians Mixture Models

Gaussian mixture models (GMMs) are the statistical and probabilistic classification approaches that employ multi-modal Gaussian distributions to capture the acoustic featured events [21]. As the variability of the feature sequences can be represented to a class through multi-dimensional Gaussian pdfs [22, 23], GMMs are currently one of the leading approaches that are used widely for the purpose of modelling and classifying tasks in audio and speech pattern recognition fields [24, 25, 23, 26, 2].

The pdf of a single Gaussian distribution can be obtained as [27]:

$$\Phi(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\Big(\frac{-(x-\mu)^2}{2\sigma^2}\Big) \tag{2.2}$$

Where μ is a mean and σ is a standard deviation. As Gaussian is the unimodal distribution, i.e. it has only one 'peak', its modelling functionality is limited to fit the points on polynomial or arbitrary distribution. Hence, the solution is to use a multimodal distribution such as mixture of Gaussians. For instance, as it can be seen in Figure 2.4, a non-symmetrical distribution with two different tails (modes) is modelled with a mixture of two Gaussians.

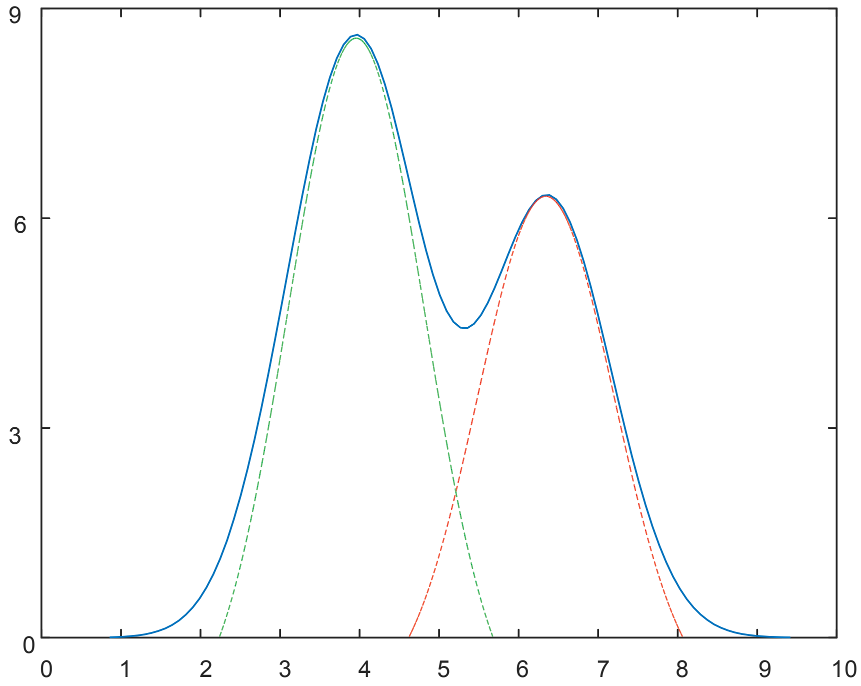


Figure 2.4: An example of a non-symmetrical distribution with a mixture of two Gaussians.

Therefore, the GMM attempts to model the distribution of feature vectors over a linear combination of Gaussian pdfs, where the mixture density of feature vector x is expressed as [27, 24]:

$$p(x|i) = \sum_{m=1}^{M} w_m b_m(x) \tag{2.3}$$

Where M is the number of mixtures; w_m is the mixture weight; and the mixture component $b_m(x)$ defines a Gaussian density function parameterised with a mean vector μ_m and a covariance matrix U_m, as [27, 28]:

$$N(x, \mu_m, \varSigma_m) = \frac{1}{(2\pi)^{D/2}|\varSigma_m|^{1/2}} exp\Big[- \frac{1}{2}(x_1 - \mu_m)\Big] \tag{2.4}$$

Where prime and D are the vector transpose and dimension of the vector x_t respectively. Generally, it is more common in speech processing tasks to use diagonal covariance matrices, as they are more computationally efficient than a full matrix, while they have the same model capability in linear GMM combination [27]. By giving an adequate number of mixtures (M), a GMM can model the arbitrary-shaped distributions [29, 23]. For instance, GMM is a powerful method for performing on the cepstral coefficients, as the cepstrum's density can be modelled by the multivariate Gaussian densities [30]. In addition, GMMs are also computationally efficient and straightforward to implement, even in real-time tasks [27].

For the training purposes, several methods are provided for estimating the GMMs parameters [31]. The maximum likelihood (ML) estimation is one of the well-established methods that employ the special case of the expectation maximization (EM) algorithm [32, 33] to obtain the GMMs parameters. The objective of the EM algorithm is to improve the parameter estimation by increasing the value of the likelihood probability that the model estimate matches the observed features, in several consecutive iterations [22].

In classification tasks, the goal is to identify the GMM model that has the maximum posteriori probability value for a given observation series. This can be done by using a likelihood function, such as the maximum a posteriori probability (MAP) method [28], which has the ability to determine the match between the parameters of the test and the trained models [30].

2.5.2 Hidden Markov Models

Basics of HMMs

A hidden Markov model is a stochastic state system with a finite number of states [34] and it can be assumed as two stochastic processes: a hidden Markov chain and an observable process through using a probabilistic function of the chain states [35]. The states in HMM can be denoted as a set of $S = \{S_1 \cdots , S_N\}$, while the transition among states, in each time index, is according to the current state transition probabilities. In other words, the transition from one state to another only depends on the current and previous states. This is known as a first-order Markov assumption [19]. In HMMs, these states are not observable, as they are "hidden". Instead, only their corresponding output or observation is discoverable. The resulting observations are assumed to be independent from each other and obtained based on the probability density function (pdf) attached to state by which they were generated.

In general, the elements of HMM can be denoted by the following parameters [36]:

- A set of states in the model $S = \{S_1, \cdots , S_N\}$, where N is the number of states in the model S.
- A set of distinct objects observed $V = \{V_1, \cdots , V_M\}$, where M is the number of

distinct observation objects.

- State transition probability distribution $A = \{a_{ij}\}$, where $\{a_{ij}\} = P(q_{t+1} = S_j | q_t = S_i)$ and $(i, j \leq N)$.

- Observation probability distribution $B = \{b_j(m)\}$, where $b_j(m) = P(O_t = V_j | q_t = S_j)$.

- The initial state distribution $\pi = \{\pi_i\}$, where $\pi_i = P(q_1 = S_i)$.

For the purpose of simplicity, as M and N describe the structure elements, a triplet $\lambda = \{A, B, \pi\}$ is the parameters' set of a compact HMM. Each various type of HMM theory can be defined by using a different state output function as the observation probability distribution. Generally, the output spectral distributions can be obtained from various modelling approaches such as: discrete, continuous and semi-continuous modelling methods [37]. Gaussian distribution [25] is one of the well-known modelling approaches that are used widely in terms of modelling the audio and speech feature vectors.

Recognized problems of HMM

The following are the three recognised problems that arise along with the HMM's implementation [36]:

1. Recognition problem: how to efficiently compute $P(O|\lambda)$ for a given model λ and observation sequence O.

2. Decoding problem: how to find a state sequence that has most likely produced O for given a model λ and sequence of observations O.

Solution: to find a maximum over all the possible state sequences, a Viterbi algorithm is employed.

3. Training Problem: how to set optimum parameters for model $\lambda = \{A, B, \pi\}$ in such way to maximize $P(O|\lambda)$.

Solution: a Baum-Welch algorithm is used to train the HMMs, which is a special case of the expectation-maximization (EM) procedure.

Recognition with HMM

The likelihood $P(O|\lambda)$ can be calculated as follows:

$$\begin{aligned} P(O, Q|\lambda) &= P(Q|\lambda)\ P(O|Q, \lambda) \\ &= \Big[(\pi_{q_1}, a_{q_1 q_2}, a_{q_2 q_3}, \cdots, a_{q_{T-1} q_T})\Big].\Big[b_{q_1}(o_1), b_{q_2}(o_2), \cdots, b_{q_T}(o_T)\Big] \end{aligned} \tag{2.5}$$

Where, $Q = \{q1, q2, \cdots, qT\}$ is the state sequence; O is the observations and the probability of the observations O and the state sequence Q given the model λ is:

$$P(O|\lambda) = \sum_{Q} P(O, Q|\lambda) = \sum_{q_1, q_2, \cdots, q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2}, b_{q_2}(o_2) a_{q_2 q_3}, \cdots, a_{q_{T-1} q_T} b_{q_T}(o_{T-1}) \tag{2.6}$$

$2TN^T$ calculations are needed to perform the above equation. Instead of these enormous calculations, there are two more efficient algorithms available as forward and backward approaches [36], for obtaining the likelihood $P(O|\lambda)$. In these two approaches the auxiliary variables called the forward and backward variables are calculated recursively, within several forward/backward iterations as follows:

Forward algorithm:

$\alpha_t(i)$ is the forward variable which represents the probability of being in the state i at the time t and observe a partial sequence $o_1, \cdots, o_t$ given the model λ as:

$$\alpha_t(i) = p(o_1, \cdots, o_t,\ q_t = s_i|\lambda) \tag{2.7}$$

If $\alpha_1(i) = \pi_i b_i o_1$, the next forward variable $(\alpha_{(t+1)}(i))$ can be obtained in each iteration t as;

$$\begin{gathered}\alpha_{t+1}(i) = \Big[\sum_{j=1}^{N} \alpha_t(j) a_{ji}\Big] b_i(o_{t+1}) \\ 1 \leq i \leq N,\ 1 \leq t \leq T-1\end{gathered} \tag{2.8}$$

Where, during the induction the $\alpha_t(i)$ is calculated at each time moment t for every state i; then, at the last step $P(O|\lambda)$ is the result by summing all the $\alpha_t(i)$.

Backward algorithm:

The backward variable, $\beta_t(i)$ represents the probability of the partial sequence that started at $(t+1), o_{t+1}, o_{t+2}, \cdots, o_T$, given the state s_i at time t and model λ as:

$$\beta_t(i) = p(o_{t+1}, \cdots, o_T|q_t = s_i, \lambda) \tag{2.9}$$

If $\beta_t(i) = 1$, the previous backward variable $\beta_t(i)$ can be obtained in each recursion step as:

$$\begin{gathered}\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(o_{t+1}) \\ 1 \leq i \leq N,\ T-1 \geq t \geq 1\end{gathered} \tag{2.10}$$

At the termination point (when the recursion reached the first state), the $P(O|\lambda)$ can be obtained as:

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1)\beta_1(i) \qquad 1 \leq i \leq N \tag{2.11}$$

The principles of the backward algorithm are the same as the forward approach; the only difference is that the backward case needs the information about the entire sequence of observations before starting its calculations. Both approaches were performed in N^2T numbers of calculation, which is much less than the calculation in equation 2.6.

Viterbi decoding

Since hidden data in HMM is represented with a set of states, the exact sequence that produces an observation sequence, is unknown. However, the sequence, which has most likely produced the given observations, can be obtained via the Viterbi algorithm in the following steps:

The highest probability along a single path, at time t, which accounts for the first t observations and ends in $states_i$, is expressed as:

$$\delta_t(i) = \max_{q_1,q_2,\cdots,q_{t-1}} P(q_1q_2\cdots q_t = s_i, o_1o_2\cdots o_t|\lambda) \tag{2.12}$$

For the purpose of finding the best state sequence (recognition step), it is necessary to find the argument which maximizes the $\delta_t(t)$; an array $\psi_t(i)$ will be used. The steps of the Viterbi algorithm are given as follows: **1- Initialisation**

$$\delta_1(i) = \pi_i b_i(o_1),\ \psi_1(i) = 0,\ 1 \le i \le N \tag{2.13}$$

2- Iteration

$$\delta_1(i) = \Big[\max_{1\le j\le N} \delta_{t-1}(j)a_{ji}\Big] b_j(o_t),\ 1 \le i \le N,\ 2 \le t \le T \tag{2.14}$$

$$\psi_t(i) = \arg\max_{1\le j\le N}[\delta_{t-1}(j)a_{ji}],\ 1 \le i \le N,\ 2 \le t \le T \tag{2.15}$$

3- Termination

$$P^*(O|\lambda) = \max_{1\le j\le N} \delta_T(j) \tag{2.16}$$

$$q_T^* = \arg\max_{1\le j\le N} \delta_T(j) \tag{2.17}$$

4- Path backtracking

$$q_T^* = \psi_{t+1}\ q_{t+1}^*,\ T-1 \ge t \ge 1 \tag{2.18}$$

The Viterbi procedure is similar to the forward algorithm, with the only difference being in replacing the summation with maximization operation over previous states. In order to avoid obtaining the extremely small probability value (in the above calculations), it is recommended [38] to use a logarithmic form of observation probability at each step as:

$$\delta_t(i) = \max_{1\le j\le N} \Big[\delta_{t-1}(i) + log(b_i(o_t))\Big] \tag{2.19}$$

HMM training

Each particular training method has a different topology structure. For instance, a left-to-right model with self-loops is commonly used in audio and speech processing [25]. The number of states in the model is chosen taking into account the modelling unit. Usually, choosing the number of states depends on the size of the modelling objects, i.e. vocalisation units in birds, word and phoneme in speech processing tasks [38].

As mentioned previously, the aim of each step in HMM training is to estimate the parameters of the model $\lambda = \{\pi, A, B\}$, where the likelihood of the training data $P(O|\lambda)$ is maximized. The task of maximizing $P(O|\lambda)$ does not have a closed-form analytical solution, hence other appropriate methods such as the Baum-Welch algorithm is used instead [36]. The Baum-Welch [39] algorithm locally maximizes $P(O|\lambda)$ by employing the ideas of the expectation maximization (EM) algorithm [32, 40, 33] to estimate the HMMs' parameters for a given set of observed features.

2.6 Clustering

Classification and clustering are two main analytic processes that are undertaken in machine learning and data mining. Generally, classification is used as a supervised learning method and it requires predefined class labels or training data, to predict a set of categorical class labels for the given test data. However, clustering is employed as an unsupervised learning method and the aim is to find a new set of partitions for the given data [41].

A clustering algorithm divides the data instances into distinct groups based on their properties of similarity. Typically, the structure clustering procedure can be described as a set of subsets $C = C_1, \cdots, C_n$, where the entire data instances or S can be formed as,

$$S = \bigcup_{i=1}^{n} C_i \tag{2.20}$$

,where, $C_i \cup C_j = \varnothing$ for $i \neq j$. In other words, any individual in the data, is associated to one and only one group [41].

2.6.1 Distance measures

To perform a clustering task, some measures between two objects need to be defined. There are two fundamental types of measures used for different clustering purposes, such as: distance measures and similarity measures [41]. Both measurement methods can determine the likeness or disparity score between any pair of individuals. Formally, the pairwise distance between two individual objects x_i and x_j is denoted by $d(x_i, x_j)$, where the corresponding value is usually non negative.

2.6.2 Similarity functions

The similarity function can be used as an alternative to the distance. A similarity function is a symmetrical function that is formally denoted as $s(x_i, x_j)$; where $s(x_i, x_j) = s(x_j, x_i)$. The pairwise similarity score is obtained by comparing the two input sequences x_i and x_j and it has a peak value when x_i and x_j are close enough to be counted as ‘similar’ [41].

2.6.3 Clustering methods

Different clustering methods can be categorised based on their induction principle. As recommended by Farley and Raftery [42], clustering methods can be divided into two primary groups: hierarchical and partitioning. As there is no precise definition available for the notion of 'cluster', many articles, including [43], suggested an alternative categorization for clustering algorithms, based on their fundamental induction [41]. As developing a hierarchical clustering algorithm is one of the main contributions of this research, thus the hierarchical methods are mainly focused on in this work.

2.6.4 Hierarchical methods

The clusters or groups in these methods are obtained by sequence partitioning the entities with several recursive iterations, in either a top-down or bottom-up procedure. Therefore, hierarchical methods can be separated as [41]:

- Agglomerative clustering - Each individual at first represents its very own group. At that point clusters are progressively merged until the desired structure is attained.
- Divisive clustering - All individuals at first are placed in one large group. At that point, the cluster is broken into sub-groups, which are progressively partitioned into their own particular sub-group. This procedure continues until the sought cluster structure is achieved.

The output of hierarchical clustering methods can be presented by a dendrogram tree. The nested clustering objects in a dendrogram, are linked with several lines, where each connecting node represents the corresponding merge or split distance, at each partitioning level (see Figure 2.5). By cutting the dendrogram horizontally, at the desired distance level, the obtained intersection points show all the formed clusters at that level.

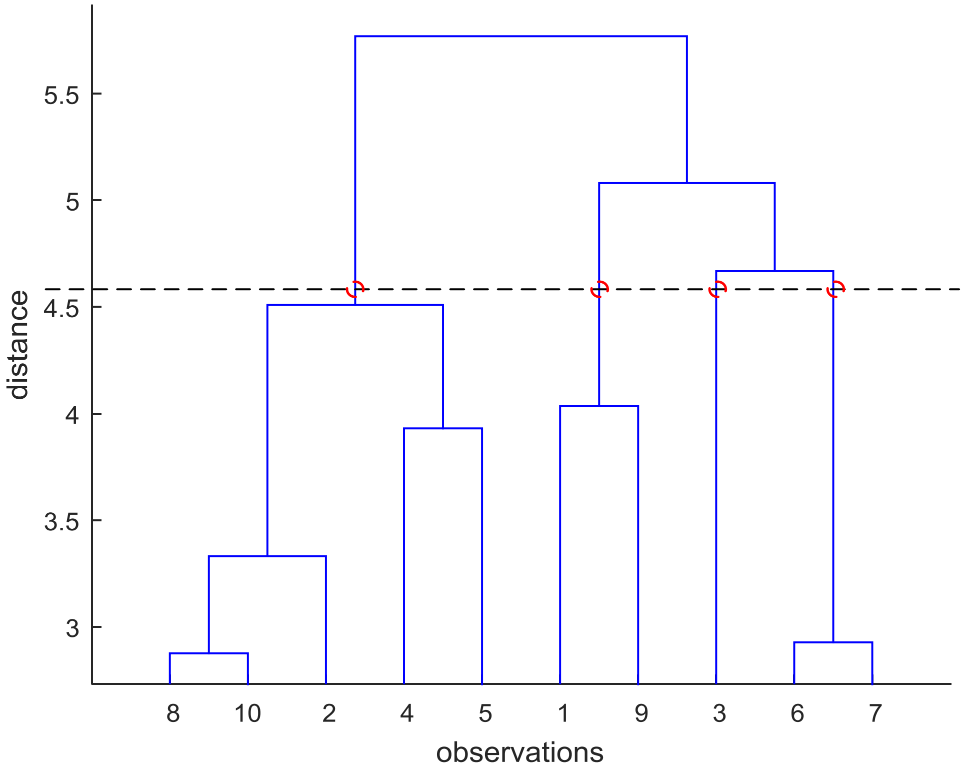


Figure 2.5: The example of a dendrogram tree for ten different observations, obtained by the agglomerative clustering method.

The merging or division of clusters is implemented by calculating the pairwise distance or similarity between the new formed group and other remaining objects or groups. There are several comparability distance measured methods available, where each can lead the whole clustering procedure to a different direction with various results (see Figure2.6). The hierarchical clustering methods could be more isolated in the way that the similarity measure is computed:

Single-link clustering (nearest neighbour): This includes techniques that consider the shortest distance between any two members of the cluster as the distance between two clusters. Assuming that the information comprises of similarity, the closeness between a pair of clusters is thought to be equivalent to the largest similarity value of any member from one cluster to any member from the other [41, 44].

Complete-link clustering (furthest neighbour): In contrast with the single-link, this method considers the distance between two groups to be equivalent to the longest distance value of any two members from the two clusters [41, 44].

Average-link clustering: In this method, the distance between two groups is estimated as the average of the distances within any object of one group to any object of the other group [41].

Figure 2.6 illustrates the different dendrograms which are obtained by employing the above distance measurement methods separately for the individual examples $a_1, \cdots, a_5$ with corresponding distance matrix D as:

$$D = \begin{bmatrix} 0 & 2 & 5 & 10 & 9 \\ 2 & 0 & 4 & 8 & 7 \\ 5 & 4 & 0 & 6 & 5 \\ 10 & 8 & 6 & 0 & 6 \\ 9 & 7 & 5 & 6 & 0 \end{bmatrix} \tag{2.21}$$

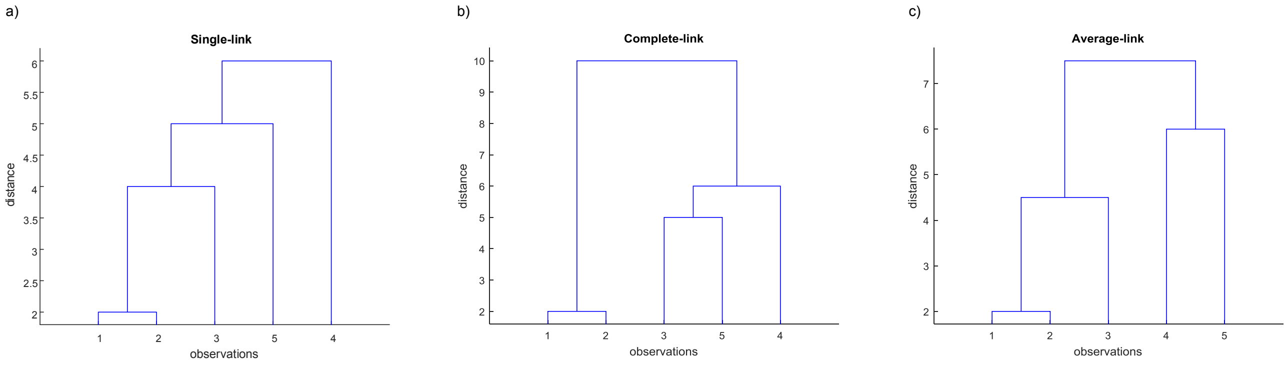


Figure 2.6: Different dendrogram trees obtained by a) single-link, b) complete-link and c) average-link clustering methods.

The main disadvantages of the hierarchical methods are:

1- The above classic methods may cause a chaining effect as: some points between two clusters can make a bridge and this allows the single-link method to merge the corresponding clusters into a larger one.

2- There is no undo or swapping procedure available in hierarchical methods; namely there is no back-tracking capability.

Having multiple partitions or divisions is the main advantage of hierarchical methods; where by defining the desired similarity level, these methods allow the users to have multiple clusters [41].

2.6.5 Standard agglomerative clustering algorithm

In general, the agglomerative method is a bottom-up hierarchical method and creates a series of divisions of the data as $P_n, P_{(n-1)}, \cdots, P_1$. At the initial stage, P_n is composed of n single-member clusters, and as the clustering ends, the last P_1 is composed of a single group of all n individuals [45]. However, in some modified versions, the clustering procedure will be stopped when the distance exceeds a pre-defined threshold. Merging at each particular stage of clustering is one of the following cases:

1- two individuals are merged into a new group of two members.

2- one individual is merged into a group containing several individuals.

3- two groups with several individuals are merged in to a new larger cluster.

As mentioned in section 2.6.4, several distance measured methods are available to define the pairwise distance (similarity score) between the new formed group and other remaining individuals or groups such as: nearest neighbour, furthest neighbour and average method that takes the minimum, maximum and the average distance among pairs of individuals within two clusters, respectively [45]. For clarity purposes, consider that there are $n = 6$ individual segments with the following similarity matrix as $SimMat_1$.

$$SimMat_1 = \begin{array}{c|cccccc|} & [1] & [2] & [3] & [4] & [5] & [6] \\ \hline [1] & & & & & & \\ [2] & 2.0 & & & & & \\ [3] & 4.0 & 5.0 & & & & \\ [4] & 6.0 & 9.0 & 3.0 & & & \\ [5] & 9.4 & 3.0 & 2.0 & 1.0 & & \\ [6] & 8.0 & 7.0 & 6.0 & 4.0 & 3.0 & \end{array} \tag{2.22}$$

At the beginning, each single segment is assumed as a distinct group $G_1, \cdots, G_6$. In the next stage, the highest similarity value in the matrix Sim_1 is that for individual 1 and 5 ($sim(5, 1) = 9.4$); as a result these two segments are merged in to a new group of two members $G_{(1,5)}$. The pairwise similarities between this new formed cluster $G_{(1,5)}$ and the other remaining individuals are obtained by using the group average method as follows:

$$sim_{(1,5),2} = mean(sim_{1,2}, sim_{5,2}) = 2.5 \tag{2.23}$$

$$sim_{(1,5),3} = mean(sim_{1,3}, sim_{5,3}) = 3.0 \tag{2.24}$$

$$sim_{(1,5),4} = mean(sim_{1,4}, sim_{5,4}) = 3.5 \tag{2.25}$$

$$sim_{(1,5),6} = mean(sim_{1,6}, sim_{5,6}) = 5.5 \tag{2.26}$$

Updating the above values in the similarity matrix gives:

$$SimMat_2 = \begin{array}{c|ccccc|} & [1,5] & [2] & [3] & [4] & [6] \\ \hline [1,5] & & & & & \\ [2] & 2.5 & & & & \\ [3] & 3.0 & 5.0 & & & \\ [4] & 3.5 & 9.0 & 3.0 & & \\ [6] & 5.5 & 7.0 & 6.0 & 4.0 & \end{array} \tag{2.27}$$

At the next step, individual segments 2 and 4 have the highest similarity score in matrix $SimMat_2$. Thus these segments, are merged into a new two-member group $G_{(2,4)}$, followed by updating the distances and the $SimMat_3$ as:

$$sim_{(2,4),(1,5)} = mean(sim_{2,1}, sim_{2,5}, sim_{4,1}, sim_{4,5}) = 3.0 \tag{2.28}$$

$$sim_{(2,4),3} = mean(sim_{2,3}, sim_{4,3}) = 4.0 \tag{2.29}$$

$$sim_{(2,4),6} = mean(sim_{2,6}, sim_{4,6}) = 5.5 \tag{2.30}$$

$$SimMat_3 = \begin{array}{c|cccc|} & [1,5] & [2,4] & [3] & [6] \\ \hline [1,5] & & & & \\ [2,4] & 3.0 & & & \\ [3] & 3.0 & 4.0 & & \\ [6] & 5.5 & 5.5 & 6.0 & \end{array} \tag{2.31}$$

This procedure of merging into larger clusters and correspondingly updating the similarity matrix is repeated recursively until the division of P_1. Finally, all the produced divisions at each stage, can be summarised in Table 2.1.

Table 2.1: Summery of clustering output at each division's level P.

Stage	Groups
P_6	G[1], G[2], G[3], G[4], G[5], G[6]
P_5	G[1,5] , G[2], G[3], G[4], G[6]
P_4	G[1,5], G[2,4], G[3], G[6]
P_3	G[1,5], G[2,4], G[3,6]
P_2	G[1,5], G[2,3,4,6]
P_1	G[1,2,3,4,5,6]

Chapter 3

THE STUDY OF BIRD VOCALISATIONS

3.1 Introduction

It is generally agreed that bird songs are among the most stunning and pleasant sounds, which exist all across our natural world. Despite the fact that everybody might expect that they can distinguish a bird melody and the way it varies from alternate sounds that other flying creatures make, the logical investigation of bird tunes has made some vital contributions to such studies as neurobiology, ethology and developmental science. As a result, this has produced an extensive and various range of literature, which can be baffling to those endeavouring to enter or study the field.

In this chapter, the scientific theories of the study of bird songs and the corresponding terminology will be introduced. Additionally, the purpose of singing and role of the songs in birds' lives will be considered and various related studies will be discussed.

3.2 Bird communication methods

Birds use various channels of communication such as acoustic, visual and olfaction (chemical). Generally, the olfactory system is not of much importance compared to the other methods as it is relatively less developed in birds [55]. Hence, acoustic and visual channels are the most notable methods that are open to most of the bird species to communicate with each other.

Birds use sound as one of their primary means of communication and there are some particular reasons behind this choice. First, sounds can travel over miles even beyond the sight of a bird, carrying information over distances despite poor climate conditions. This gives the birds the ability to communicate with their flock and make their presence or their place known to them. For humans as well, acoustic communication is one of the main ways of detecting the birds in conditions where they cannot be easily seen. Secondly, sound is a rapid way of transmitting the information efficiently. Birds produce songs based on the demand, so songs carry a vast amount of information [56, 57].

Different species of birds produce different types of songs, which make these sounds even more interesting to humans as they can use them as a way of distinguishing the species of birds and the diversity of the regions of study. Although birds produce different types of sounds such as songs, calls, warbles, trills, croak, drums and whistles, birds' acoustic sounds can generally be divided into two main categories: vocal and mechanical (non-vocal) sounds. The majority of bird species use their own vocal organs, such as the syrinx, to generate different types of acoustic sounds. The latter refers to species-specific audio signals that are not produced in the bird's syrinx.

However, some bird species generate non-vocal sounds by movements of particular body parts, such as wings and tails, or anatomically adapted features. For example, ruffed grouse birds make a non-vocal sound by flapping the air with their wings. The movement of the wings creates a vacuum and air barrels through this space and creates a mini sonic boom [58]. In other examples, woodpeckers produce a drumming sound by pecking on a resonant object and pigeons and doves use their wings to create clapping sounds [59]. The main focus in this study is merely on sounds that are produced by a bird's vocal organ and the term 'acoustic sound' refers to vocal sounds of birds.

As birds have quite sharp sight, visual demonstration is also an important means of communication for them. These signals are mostly used in daytime as their sight is low in the darkness of the night and also, the same as humans, birds are active and awake creatures during the day. The form of signals they use may vary from colours (like bird feature) to body postures.

In birds, these visual signals are the product of certain postures in the body and the features of the bird [58]. They use visuals signals mostly when they are seeking mates or defending their territory. They may use and combine both acoustic and visual signals to make their statement more clear.

Visual signals are of an advantage where using other signals would not be safe. However, they have poor functionality in many particular situations such as at night, in misty weather and darkness habitats. On occasions like in dense habitats (forests, reeds) where birds can easily move out of sight behind other objects, using visual signalling cannot be an efficient way of communication [55]. On the contrary, sound signals can travel to any corner, pierce through the objects or move around them and reach beyond the miles.

3.3 Singing behaviours

In this section, the singing behaviour of birds is summarised, including the context of songs, the occasion of their occurrence and the producer in particular.

3.3.1 Male singing behaviours

Functions of birdsong are being actively studied in behavioural ecology research [60]. In general, the male singing behaviours can be divided into two functional manners:

1-Territory defending or male-male vocal fighting:

Males of the various species use songs as the first step of their mate fighting [55]. Vocal fighting may happen in two aspects; taking over a new breeding site or defending the current territories by producing repellent signals. In general, male species may contest over mates or territory, such as a nest or feeding site, as well as being a means of attracting females [60].

Several naturalists like Gilbert White [61] and Eliot Howard [62] have observed a relationship between birds singing and their territory as a fundamental necessity to draw the attention of females. For a male owner of a territory, singing is a way to brag about his power in fighting and imply to its rival that in the case of battle he will definitely be the loser. Therefore, birds use songs to make their rivals withdraw from the combat and then both of them will benefit [60].

Therefore, the song's characteristic should reflect the fighting motivation as well as assets that can lead to winning of a battle; factors such as physical strength (size, weight and body conditions) and combat skills can play an important role in a rivalry competition between males [60].

2- Attracting mates or female choice:

With birds, it is usually the responsibility of males to attract females; in general it is only the males of the species who sing [60]. According to the literature [55, 60], for birds, songs are also the principle way of appealing to other mates and showing off their overall health. For females, reproductive accomplishment is the main criterion in finding a mate; so characteristics like age, condition, parental ability and the quality of his territory is of importance when it comes to choosing their partner. Other males are also interested in knowing of their rivals, their location, the possibility of an attack from them and their fighting skills [60]. Although sometimes both males and females will find information of their interest in the same factors in a song, in most cases this information is quite different and will be extracted from various song aspects [60].

One simple experience shows that if a song is broadcasted from loudspeakers in a territory which is not protected by any male bird, the territory will remain safe from intruders for a longer time compared to when it is in silence [63, 64, 65]. The broadcasting song can attract females to that territory as well [66, 67, 68]. It has even been observed that females perform copulation solicitation displays in response to the song and get close to the speakers [69, 70, 55].

3.3.2 Female singing behaviours

Although it is largely believed that singing is under the control of male species and females are generally harder to observe singing [55, 60, 71], many studies have found that the female, in some bird species, may also sing. For instance, singing has been observed with female superb fairy-wrens [72], white-crowned sparrows [73], blue-breasted waxbills [74], European robins [75] and long-tailed manakins [76]. Generally, female birds sing regularly during autumn, winter and spring prior to breeding and their songs are structurally similar to male songs, but they are shorter and simpler [55, 77].

Females may sing either solo or in a duet, for broadly the same reasons as males. Their singing may have similar functions as well. One main function of singing for a female is to attract a male partner. They also may sing to defend their territory from possible intruders. For example in fairy-wrens, as the males are often away from the territory, the females take over the singing duty to defend their territory, although at the same time, their singing may attract other males to their territory to engage in extra-pair copulations [60]. In robins and mockingbirds, when the food resources are low in winter, the females sing to defend their territory from intruders who come looking for food [75]. In several polygynous species, singing is a means of aggression of one female to another who wants to engage in extra-pair copulations with her mate [78].

In parallel with common research on female singing behaviours, in a recent study [79], the authors investigated over 200 European passerine species to see whether the females produce songs. Based on their findings, they concluded that the importance of female singing in birds is broadly underestimated [79]; and they suggested that there is a need for focused studies on female vocalization to get a deeper understanding of their interesting behaviour.

In conclusion, in most species, females have been observed to use songs for quite the same functions as the males, such as territorial defence and mate attraction. Females also participate in producing duets with males which will be discussed later in this chapter [55].

3.3.3 Early morning singing behaviours (The dawn chorus)

Studies on bird songs all reveal that the best time of the day to observe, record and perform field experiments is during the early morning. The birds start singing well before the sunrise, even before the sky begins to lighten and they will sing for two-three hours. The peak active time of singing in male birds is about dawn, when a group of birds from many species may sing together in what has become known as the dawn chorus [55].

Although most species participate in this dawn chorus, different species start singing at slightly different times. For instance, observations in an English woodland reveals that robins, blackbirds and song thrushes start singing earlier, known as ‘early birds’ for that reason, while chaffinches and blue tits start later [55].

So, why is the early morning the best time for singing? Many early and recent studies such as, [80, 81, 82, 83], reveal that their starting time of singing depends on their seeing abilities in the low light situation of the early morning; this is why the early birds have been noticed to usually have larger eyes. In 2002, Thomas et al. [82] studied the relationship between the size of the eyes and the diameter of the pupils of birds with the time they start to sing; this study confirms this hypothesis. Berg et al. [83] carried out wide studies on tropical species and also revealed a similar conclusion regarding the correlation between a species’ eye size and the time at which they began to sing. However, foraging height was the most effective variable predictor with canopy species, as they start to sing earlier than the species that live at the lower levels close to the forest floor. In general, there are three main benefits of singing at dawn: it is a good condition for sound transmission; there are low visibility conditions and overnight abandoned territories [55].

3.3.4 Duet singing

In general, duetting is when birds sing with one another, either at the same time or in turns [71]. Many studies have observed duet singing in birds and this section is a brief review of some of the relevant works.

In some species like the dusky antbird [84] or the Polynesian megapode [85], a duet may be used as a more effective way of a territory rather than a solitary defence. In other species like bay wrens, the duet may be used for mate guarding [86, 87], but in this case the role of a male counterpart of a duet is different from the role of the female counterpart. The author [86, 87], found that the female sings to confront her same-sex intruders, whereas the male sings to protect his mate from extra-pair copulations.

In long-tailed manakins, two males engage in singing as a pair to attract a female, one of which is usually the leader and the other is the follower. They may sing together for a long time and match the frequency of their songs [76].

3.4 Bird vocalisation structure

Bird vocalisations are generally categorised as songs and calls, by the context of their length and the functionality [55, 81].

Songs are usually long and complex. These controlled vocalisations are generally produced by males to attract females or to defend their territory and they mostly occur in the breeding time of the year [55].

Calls are shorter and simpler. They are not supposedly rhythmic, may be produced by both males and females and also occur all through the year. Calls usually carry information like fight, threat, alarm, etc [55].

Although this division is a little unscientific and more traditional, the above terms needed to be briefly discussed as they are still in use. To achieve a deeper analytical classification, it is crucial to inspect acoustic signals visually.

3.4.1 Sound visualisation

Before the sonogram became widely used in bird sound studies in about the 1950s, our ears were the only tools to study birdsong. Until that moment, the cathode-ray oscillograph machine was the only available device which was used to visualise the sound waves. The oscillograph machine displays the given sound signal as a two-dimensional graph, in which the x-axis shows the time index and the y-axis shows the sound wave pressure. However oscillograms are only helpful to study sounds of insects and not for analyzing sounds which have a complex frequency structure, like birdsongs or human speech [60].

By developing the first generation of sonogram machines around 1940-50, the use of speech sound spectrograms became a landmark in linguistics research. Later, in about 1953, Donald Borror [88] and Nicholas Collias [89] were the pioneers who employed spectrograms in their early birdsong studies. From that time until now, spectrograms have served as a powerful method for understanding and analyzing the description of bird vocalization signals [60].

As it can be seen in figure 3.1, spectrograms are basically a graphical representation of the loudness or the amplitude of an acoustic signal (birdsong) in different frequencies over time. A spectrogram has two dimensions of time and frequency and a third dimension of magnitude which is represented in colours. The horizontal axis is time, going from oldest to youngest and the vertical axis is frequency or pitch going from low to high. The loudness or amplitude is represented on the third axis by colours, going from dark blue (low amplitude) to red (strong amplitude).

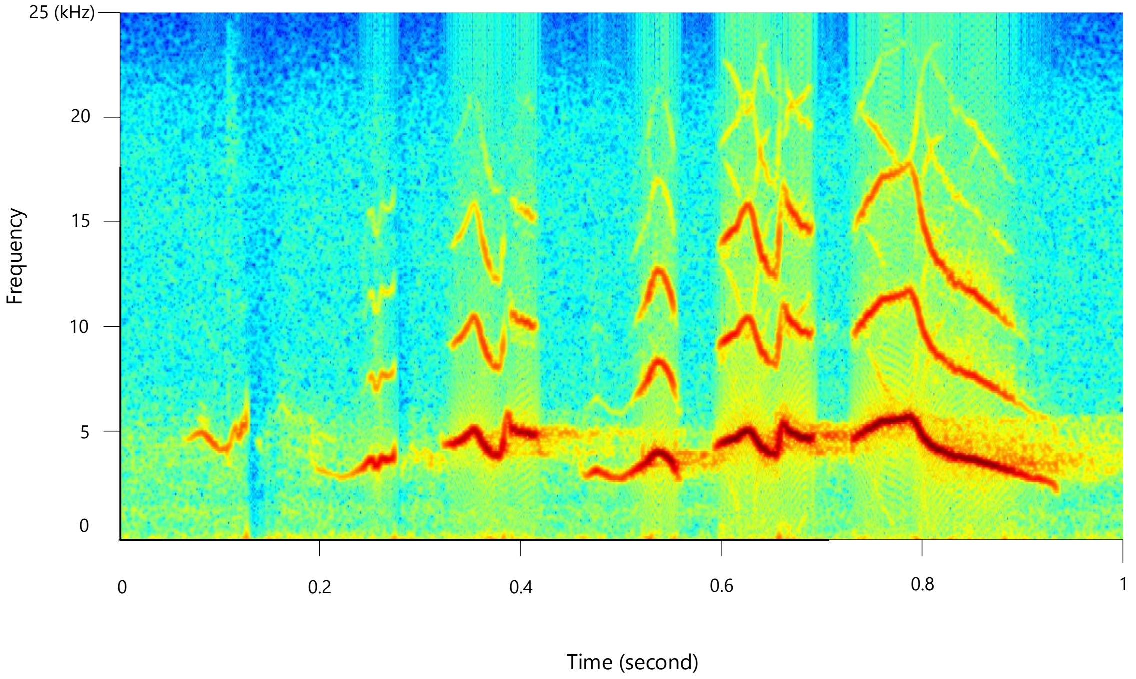


Figure 3.1: An example of a spectrogram of a birdsong (Magnolia Warbler)

3.4.2 Hierarchical terminology of bird song

As complexity is one of the main differentiating parameters that separate most songs from calls, the development of sound spectrography was a turning point in studying bird vocalisations. The spectrogram provides discriminating means to define various structures in a birdsong [90]. By this means, each song can be divided into hierarchical levels of phrases, syllables and elements. As it can be seen in figure 3.2, elements or notes are the smallest units in a bird's song' being defined as a single line in the sonogram. Syllables comprise of one or several elements, depending on their complexity. Phrases are a series of syllables that occur in a pattern [55].

Another factor in defining these structures is their time intervals, with songs having the longest intervals and elements having the shortest intervals and calls and syllables are in between. There are great variations in the forms and structures in songs which make it hard to reach a fixed definition for these terms, as one may face slightly different descriptions across the research literature. However, for the purpose of this thesis, the above definitions will be used as a general guide [55].

3.4.3 Song type

For the purpose of machine learning (not a biological perspective), in terms of syllable complexity the birds' songs can be divided into four different categories as follows [91]:

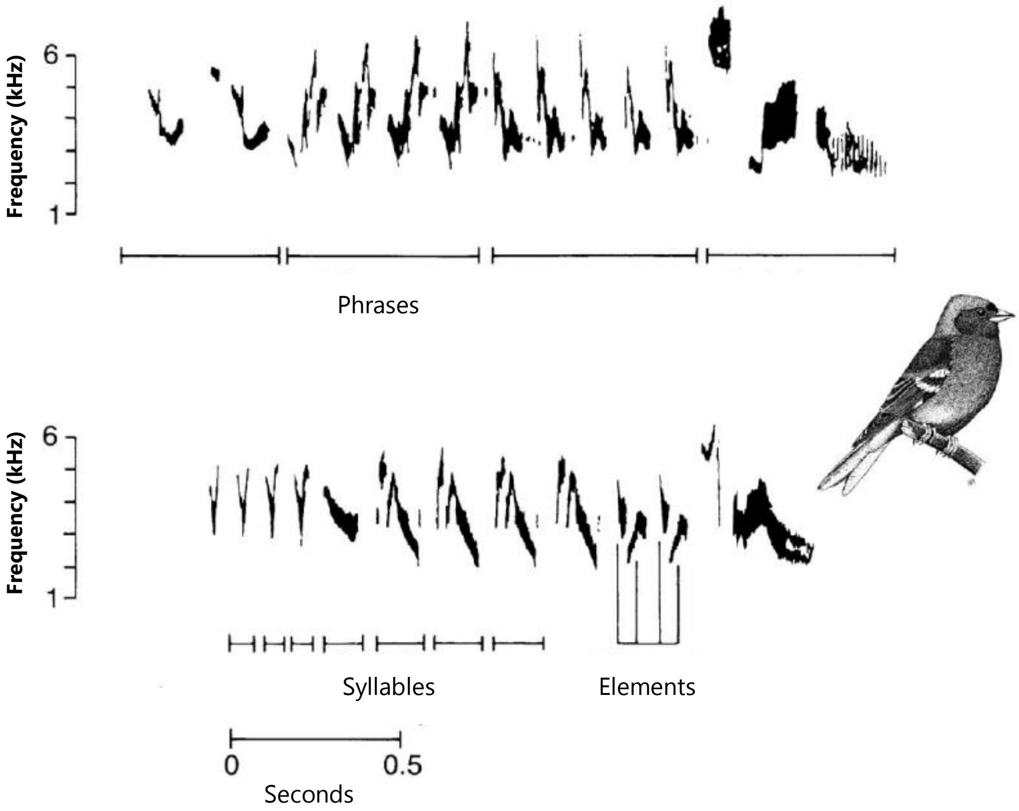


Figure 3.2: Hierarchical divisions of two different songs of a male chaffinch [55]

Monosyllabic:

Most of the bird calls are short and monosyllabic. However, in some birds e.g. sparrowhawk and green woodpecker [91], the song is only made of a single syllable. Typically, these monosyllabic songs will be repeated sequentially during a short vocalisation.

Multi syllables:

In this type of vocalisation, the song is composed of more than one syllable, typically two or three. These types of songs can be sung by the majority of passerine birds such as:

wrens, tits and sparrows.

Large vocabulary:

For a small percentage of bird species, even in the UK, their songs have a complex structure and may consist of some random syllable sequences followed by some fixed patterns. For example, the skylark employs about 300-400 different vocabularies within her vocalisations [91, 92].

Less tonal:

The song structure in some bird species, e.g. jays, crows and screeches, is less harmonic or tonal. This type of song does not preserve a specific energy within the signal and the frequency range varies about 1-4 kHz.

3.5 Sound production

As mentioned previously, birds can produce an astonishing variety of songs that often remind us of the spiritual springtime. However, there is only one order of birds who sing, namely the passeriformes, which comprises of only half of the birds on the planet [93]. The rest merely use calls to communicate. The passeriformes mostly generate complicated vocalisations or songs, thus they became known as 'the true songbirds' [55]. Overall the principal question is how can birds make these inventive and elaborate vocal sounds? To answer this question, a brief survey on birds' vocalisation systems is essential.

The song production journey in birds starts in their brain, where the distinct neural pathways continuously control the vocalisation procedure [59, 94]. However, for the sake of simplicity, this section briefly describes the principle process of the birds' vocalisation mechanism by only considering the role of their vocal organs rather than their brain's pathway.

3.5.1 Vocal tract's modulations

Many studies [95, 96, 97] confirm that birdsongs have a pleasant tonal structure. In birds, this tonal feature is attained by the creation of non-harmonic pure sounds within a restricted frequency range [55]. The question is does the sound source in birds produce this tonal quality of sound or does the bird's vocal tract contribute to such beautiful singing?

Human speech and birdsongs have similar qualities when considering the vocal articulation [98]. In both songbirds and humans, the vocalisation sounds are produced by the flow of air through a vocal system. In humans, the exhalation of breath from the lungs generates a complex waveform at the vocal folds; the components of this waveform are subsequently modified by the rest of the vocal tract including the mouth, tongue, teeth, and lips [98]. The human's vocal tract acts as a filter by creating concentrations of energy at particular frequencies, called formant frequencies. As an example, vowels are characterized by relatively constant formant frequencies over time; whereas during consonant production, the formant frequencies change rapidly (20-100 ms), resulting in formant transitions [98].

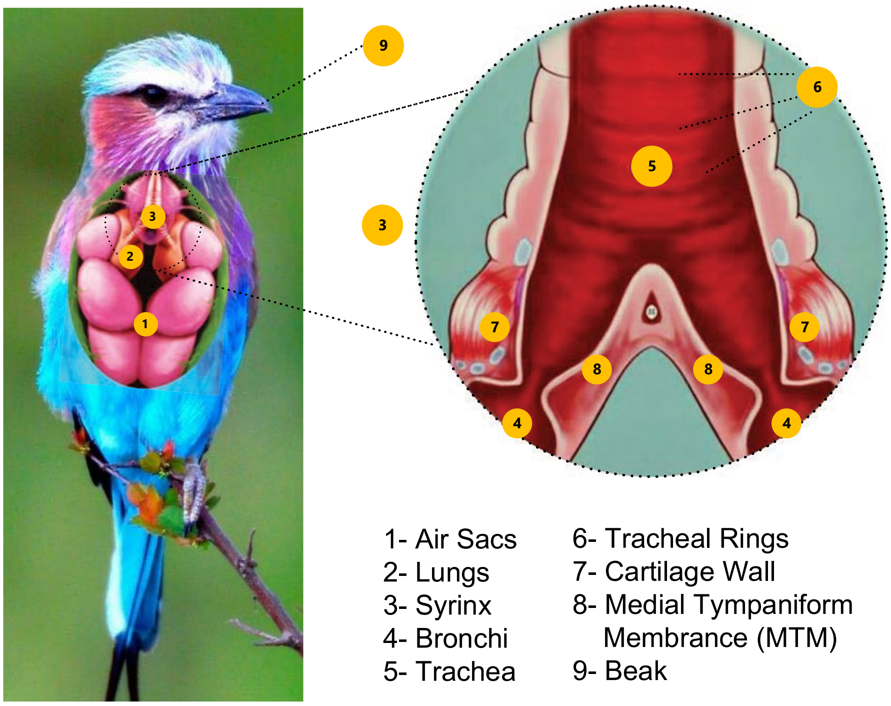


Figure 3.3: Vocal organs of a songbird

As figure 3.3, the main organs of the sound production mechanism in birds are the lungs (as in humans), the syrinx (with analogies to the larynx in humans), the trachea, the mouth and the beak (acts as the nasal cavities in humans) [98, 12, 99]. In songbirds, sounds are generated by the flow of air during expiration through the syrinx. The syrinx is the most important organ in the sound production mechanism and it is the bird's voice box, a bilateral structure surrounded by specific muscles. Unlike the human larynx, which is at the top of the trachea, the syrinx is situated lower in the bird's chest at the end of the trachea (acts as a resonator) and on top of the two bronchi. The airflow from the lungs to the top of each bronchi, makes a syringeal medial tympaniform membrane (MTM) to vibrate nonlinearly opposite to the cartilage wall. The bird can control the pitch by changing the tension on the MTM and can control both the pitch and volume by ranging the force of its exhalation [12]. The syrinx's different position surely contributes to the complexity of the songs in birds, as it benefits from two bronchi for two potential sources of sound [98]. This makes it possible for birds to produce two simultaneous notes and pitches by controlling the two branches of the trachea independently [59, 99].

Furthermore, just like humans, birds use their upper vocal tract as a selective acoustic filter to modify the frequencies in the final sound. In general, the sound velocity and vocal tract cavity determine the acoustic resonance in both humans' and birds' vocalisations procedure [55]. As humans can adjust their speech's resonant properties merely by changing the position of their tongue, jaw, lips and the buccal cavity [55, 100], birds also can control their vocal filter resonances through a variety of actions; such as stretching or retracting their neck and expanding their throat and beak movements, which is the main factor in many species [101, 102, 55, 103]. Westneat et al. (1993) [101] and Hoese et al. (2000) [102], found apparent connections between the width of the beak opening (gape) and the produced sound frequency by studying head and beak motions in singing birds. Hoese et al. (2000) [102] stated that birds mostly open their beak widely to generate a higher frequency sound and close it more for a lower frequency sound. Accordingly, for the production of a wide frequency syllable, like a trill sound, the bird's beak should perform at a broad angle [55]. In terms of how fast this action can be done (performance rate), there is a trade-off correlation between bandwidth and repetition rate [55]. Hence, in a study [104], by setting some experiments on American sparrows, the author discovered that the wide bandwidth (high-pitched) sounds can be only generated at a slow repetition rate; whereas the narrower bandwidth (low-pitched) sounds can be performed at any repetition rate.

The final generated birdsong may consist of components which are purely sinusoidal, harmonic, non-harmonic, broadband and noisy in structure. In general, amplitude modulations of the fundamental element are mostly produced by the syrinx, but intensity differences between harmonics is based on the properties of the upper vocal tract [12].

As discussed previously in section 3.4.2, by visualising the spectrogram, a bird vocalisation song can be divided into a set of syllables. As can be seen in Figure 3.4, (in terms of pitch pattern classification), all bird vocalisation elements can be described by five basic patterns as follows [105]:

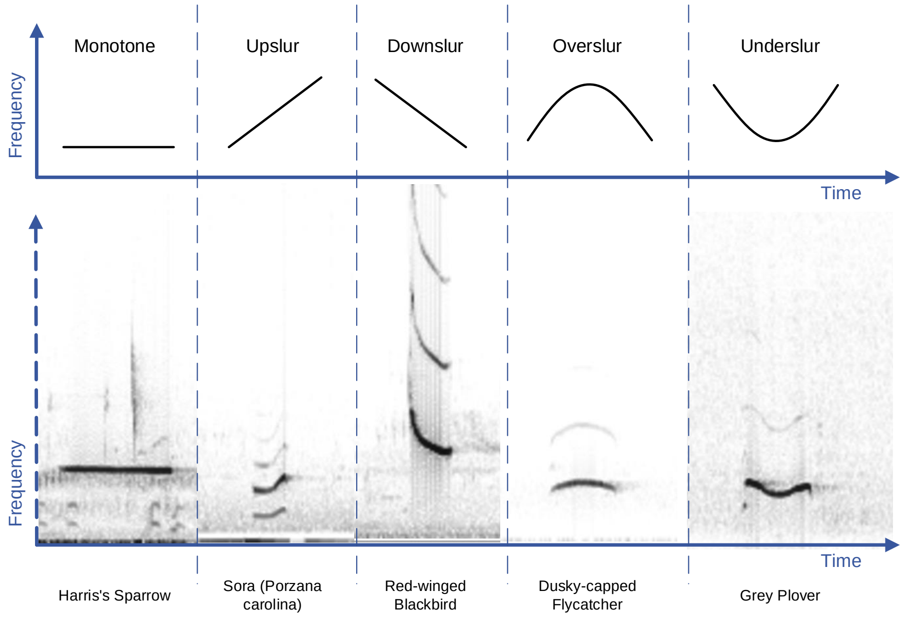


Figure 3.4: Five basic pitch patterns with real examples

1- Monotone: the sound does not go up or down, just remains at the same pitch from start to end.

2- Upslurred (or rising): the sounds rise in pitch and appear tilted upwards.

3- Downslurred (or falling): the sound falls in pitch and appears tilted downwards.

4- Overslurred: the sound rises and then falls in pitch, appearing and sounding highest in the middle.

5- Underslurred: the sound falls and then rises, appearing and sounding lowest in the middle.

Moreover, each of the above defined categories can be expanded into the others' subgroups, separately in terms of length and the frequency boundaries. In other words, these vocalisation patterns can be found to be different, in terms of length and the frequency boundaries, within different bird species. Later in Chapter 6, the diversity of these vocalisation patterns for each bird species will be discovered (see Figure 6.25).

Typically, birdsongs can produce a large set of different sounds in frequency ranges between 100 Hz and 8 kHz, but the variation between different species is large [12]. In addition, this variation of frequency boundaries may be found within an individual species in different geological locations or seasons. The principal question is, how could this range variation happen in bird vocalisations' signals?

Body size and anatomy variation over the different bird species: in addition to the vocal tract's modulation roles, there is a relation between the frequencies that birds are able to generate and their body size. The larger the body size, the lower the frequency of the bird's sounds and vice versa [55, 106]. Furthermore, the distinction between anatomy and the size of the sound production organs (e.g. syrinx and trachea) in different species, can affect the frequency range and the pitch information on vocalisation sounds. According to [12], three different types of syrinx, namely tracheobronchial, tracheal and bronchial, can be found among bird species because of the distinction between the tracheal and bronchial elements of the syrinx and the topographical position of the main sound producing mechanism. Furthermore, the number of the tracheal cartilage rings, which act as a resonator to the produced sounds, depends on the length of the neck; it ranges from about 30 in small passerines to about 350 in long necked flamingos and cranes [12].

Local dialects: just as humans have regional accents, some bird species develop distinct, area-specific dialects. Such variation in song often arises when populations of the same species are isolated by geographic features such as mountains, bodies of water, or stretches of unsuitable habitat. These local dialects are then passed on to the next generation of young birds, which hear the songs being performed by their father and other local males. After many generations, the birds from one area can sound quite different from those from the next mountain over.

Brain hormones and seasonal songs: birdsong is seasonal and as song production is controlled via a pathway beginning in the brain, existence of circulating hormones, e.g. testosterone, play an important role in song production over the year. According to [60], experiments have established that the vocalisation frequencies of the same pattern-structured song, are changed up to 400 Hz over the year for some bird species. In the next section, we will discuss more about the learning and development of birdsong.

3.5.2 Vocal learning and Development of bird song

In the Passeriformes bird species, vocal learning performs an important role in their singing abilities. Many early studies, such as [88, 107, 108, 109, 110], found that some species-specific characteristics of bird song can be obtained by listening to other individual birds. In the late 1950s, William Thorpe [107] demonstrated evidence on the influences of vocal learning by studying two different groups of young chaffinches [107]. He kept and raised the first group in an isolated acoustic laboratory, where the recorded songs of an adult wild chaffinch were played. In contrast, in the second group the imitation songs were removed. The birds who were in the first group sang species-specific song pieces as adults; contrariwise, the second group of birds sang anomalous songs. These results showed that the birds must learn how to sing by listening to their tutor in the early seasons of their life. Moreover, in 1970 Peter Marler [110], by introducing a comparative approach between vocal learning in white-crowned sparrows and language learning by humans (speech developments in children), found that birds have an inherent ability to learn the song dialect during a sensitive period.

In birds, learning to sing takes some time to complete. Typically it begins in the first month after their birth and continues for few months, until just before the bird sets up his own territory. However, in some bird species this takes longer than in others and also they may learn additional songs within the first two years of their life [60]. Generally, similar to human speech development, vocal learning and sound development procedures in songbirds can be divided into two main stages: the sensory phase and sensorimotor phase [111, 112].

Sensory phase:

During the sensory or memorisation phase, young birds train their sensory system by listening to the adults' singing. Thus, this phase starts by memorising the songs that are produced by an adult tutor [111, 112, 113]. During the first year of life, though it may vary among different species, the young bird learns the species-specific sounds.

So how do young birds recognise the same species' sounds over multiple birds' species that are living in their neighbourhood without having advanced hearing experiences? Some recent studies in the ornithology field, such as [114], demonstrate that birds are often born with an innate mental power to distinguish their own species-specific songs from other bird species. Along with this genetic predisposition, the specific timing of the sensory phase within species [111], may increase the chance of learning the correct species-specific songs.

In addition, a recent study [91] named the above-mentioned locality learning effect as the ‘regional accent effect’. The study then claimed that the existence of common syllable vocalisations within individual species that sing in the same neighbourhood may refer back to this effect.

Sensorimotor phase:

In this phase, the young birds start making motor patterns out of this sensory memory and producing songs [113]. This means that over time and by practicing and comparing their vocalization to the template, they will learn to sing [111]. Also hearing in this phase, plays an important role in birds’ singing behaviours. Konishi in his study [115] on the effects of hearing their own vocalisations in a bird’s song development, found that even if a bird deafens after the sensory phase and before going through the sensorimotor phase, she still produces some abnormal kinds of songs [115].

In general, in this stage the birds generate three different types of songs (in terms of the structure and the loudness) such as: subsong, plastic and crystal. The first type is named ‘subsong’ and it is generated during the initial stage of the sensorimotor phase. These songs are almost silent, roughly structured and very unstable in form. As they go further, the songs get louder and improve in structure, but they still do not have consistent forms. Eventually, the structure of these plastic songs will become quite similar to the version they have been heard to sing during the sensory model. Finally, by the existence of circulating testosterone (T) [116, 117] in birds, they can produce crystal songs which are the stereotyped form of the sensory model [113].

3.6 Conclusion

This chapter has discussed the basic biological theories of bird vocalisation. It began with studying the communicational behaviours of birds, followed by an explanation of their singing behaviours. Then, by using the spectrograms, the birds' vocal signals were distinguished into the various hierarchal levels of bird sound, such as songs, calls, syllables, phrases and elements. In the sound production section, the mechanisms of sound production in typical birdsongs were discussed by describing their vocal organs. Finally, the sound development procedure in birds was studied with two various learning phases, the sensory and sensorimotor phase.

Chapter 4

BIRD VOCALISATION CORPUS

4.1 Introduction

The classifying of sounds of birds based on their species has been the subject of many studies in the last two decades, dating back to McIlraith and Card [10] in 1997. In the early studies, the databases which were used were quite small, often free of noises and/or being manually segmented. Furthermore, the number of species the studies studied was relatively small, which made them impractical in ecological applications. As a result, there developed an increasing demand for a large database of recorded birds' sounds for research purposes [118]. In other words, the amount of training data which is used to build the training models, is an important factor for assessing the performance of different classification approaches. Thus the availability of a large dataset of bird vocalisation sounds is crucial.

In the following sections, first a brief literature review of the large, available bird vocalisation archives is presented, followed by a description of the data which has been used in recent bird classification challenges; then a large database of bird sounds has been classified and labelled into different sound events manually.

4.2 Available large-scale bird sounds' archives

Only a few bird sound archives exist in the ornithology field which provide a large collection of data sets of bird sounds. Some of these are laboratories or websites, such as xeno-canto [119]; and the websites are community databases which share their entire collection publicly on the Internet. Other archives, such as the Borror Laboratory of Bioacoustics [120] and the Macaulay Library [121], are only open partly to the public and the rest, including high-quality copies of recordings, are accessible to researchers and other professionals merely through their requesting or ordering procedures.

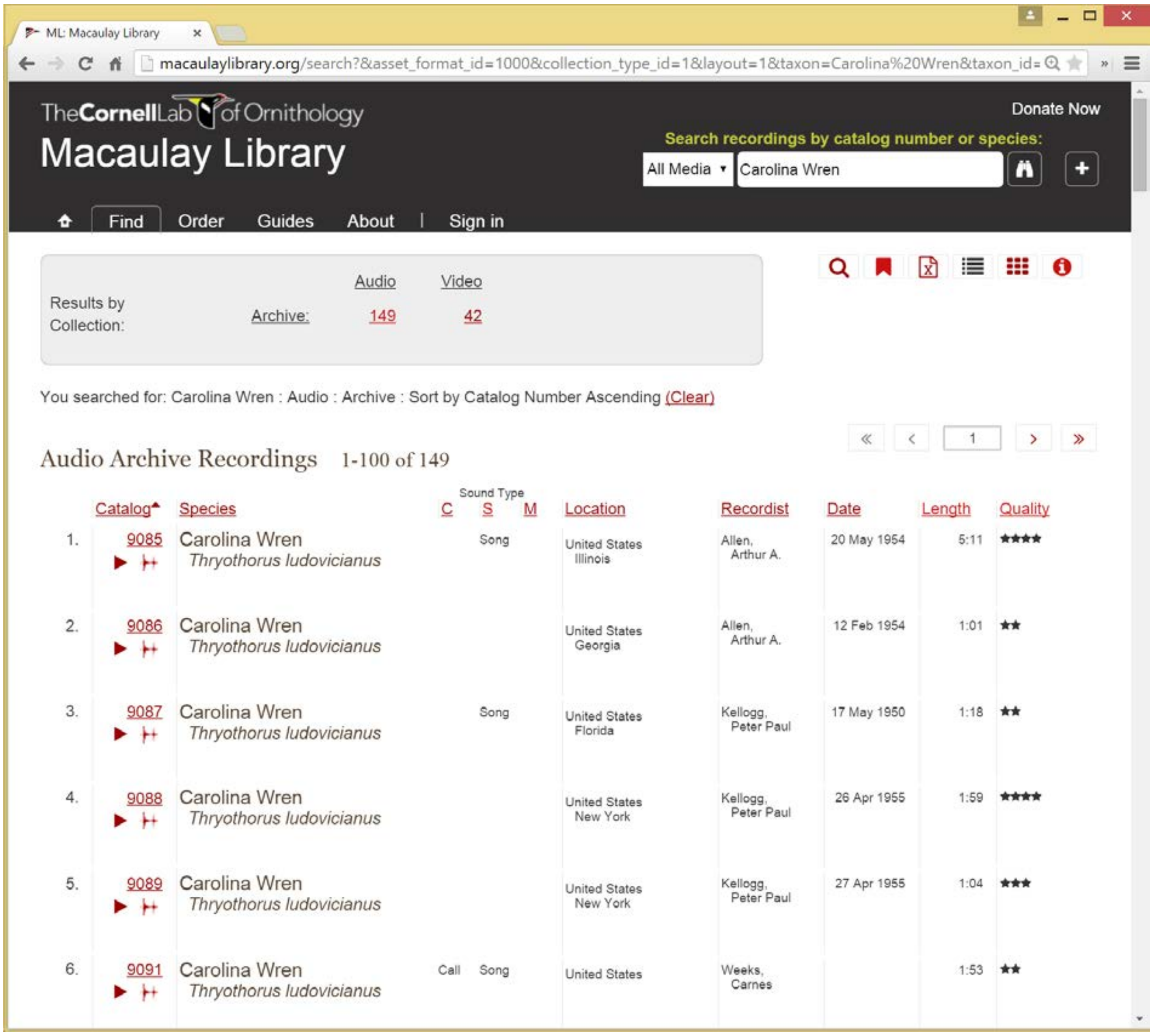
The following is a brief review of three different accessible archives which are used in related bird sound identification or classification research. As this data has usually been recorded in natural fields, the files are commonly infected with several background sounds other than bird vocalization, such as the sound of wind or water, speech and other noises. However, all the recorded files provided by the libraries or archives below, have been merely labelled with the name of the bird species, with no further segment level annotation information available with them.

4.2.1 The Macaulay Library

The Macaulay Library is part of Cornell Lab of Ornithology [122] and is the world's leading scientific collection of biodiversity media. According to the information provided on their website, their archive with more than 175,000 audio and 60,000 video recordings provides a vast documentation on the diversity of behaviour among birds and other animals. The library includes more than 130,000 recordings of bird sounds, more than 70% of which are Passeriformes or tonal bird species from all over the world (though mostly in the US). There is a small description note available with each recording representing the recording quality level (score from one to five stars), bird sounds behaviour, date and the location of the recordings (see Figure 4.1).

4.2.2 Xeno-Canto

Being a community database, xeno-Canto has provided a worldwide collection of recordings of wild bird sounds on its website. Their archive comprises more than 224,000 recording files from over 10,000 different bird species, including more than 128 different Passeriformes. The entire database can be searched based on different parameters such as the bird's name, time and length of the recording, location and the recording quality levels (see figure 4.2). The recording quality of each sound file is described with alphabetic label characters from A to E (the best and worst quality are labelled with characters 'A' and 'E' respectively).


ML: Macaulay Library
macaulaylibrary.org/search?&asset_format_id=1000&collection_type_id=1&layout=1&taxon=Carolina%20Wren&taxon_id=
TheCornellLab of Ornithology
Macaulay Library
Donate Now
Find
Order
Guides
About
Sign in

Search recordings by catalog number or species:

All Media: Carolina Wren

Results by Collection:		Audio	Video
	Archive:	149	42

You searched for: Carolina Wren : Audio : Archive : Sort by Catalog Number Ascending (Clear)

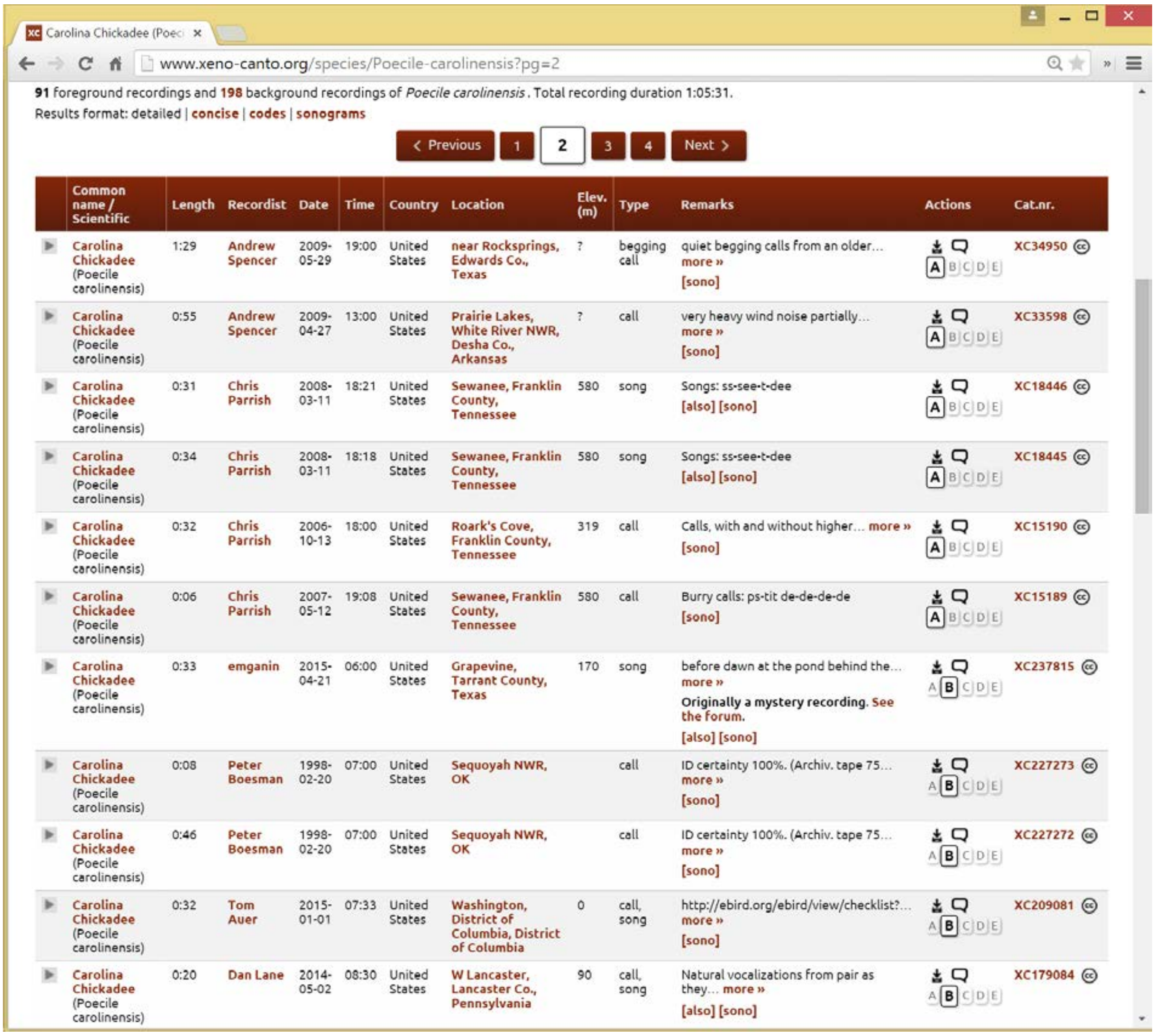
Audio Archive Recordings 1-100 of 149

	Catalog	Species	Sound Type C	S	M	Location	Recordist	Date	Length	Quality
1.	9085	Carolina Wren *Thryothorus ludovicianus*		Song		United States Illinois	Allen, Arthur A.	20 May 1954	5:11	★★★★
2.	9086	Carolina Wren *Thryothorus ludovicianus*				United States Georgia	Allen, Arthur A.	12 Feb 1954	1:01	★★
3.	9087	Carolina Wren *Thryothorus ludovicianus*		Song		United States Florida	Kellogg, Peter Paul	17 May 1950	1:18	★★
4.	9088	Carolina Wren *Thryothorus ludovicianus*				United States New York	Kellogg, Peter Paul	26 Apr 1955	1:59	★★★★
5.	9089	Carolina Wren *Thryothorus ludovicianus*				United States New York	Kellogg, Peter Paul	27 Apr 1955	1:04	★★★
6.	9091	Carolina Wren *Thryothorus ludovicianus*	Call	Song		United States	Weeks, Carnes		1:53	★★

Figure 4.1: Example of information provided along with recording files in Macaulay library [121].

4.2.3 Borror Laboratory of Bioacoustics

As one of the earliest and most comprehensive databases of animal sound recordings, this still-growing archive comprises over 40,000 recordings of animal sounds, in company with more than 700 audio files of different Passeriforme bird species. All these data were recorded by field observers of the natural habitats of birds, mostly in aspen forests, dense forests and marshlands of the western United States over past decades.



Carolina Chickadee (Poeci ×
www.xeno-canto.org/species/Poecile-carolinensis?pg=2
< Previous
1
2
3
4
Next >

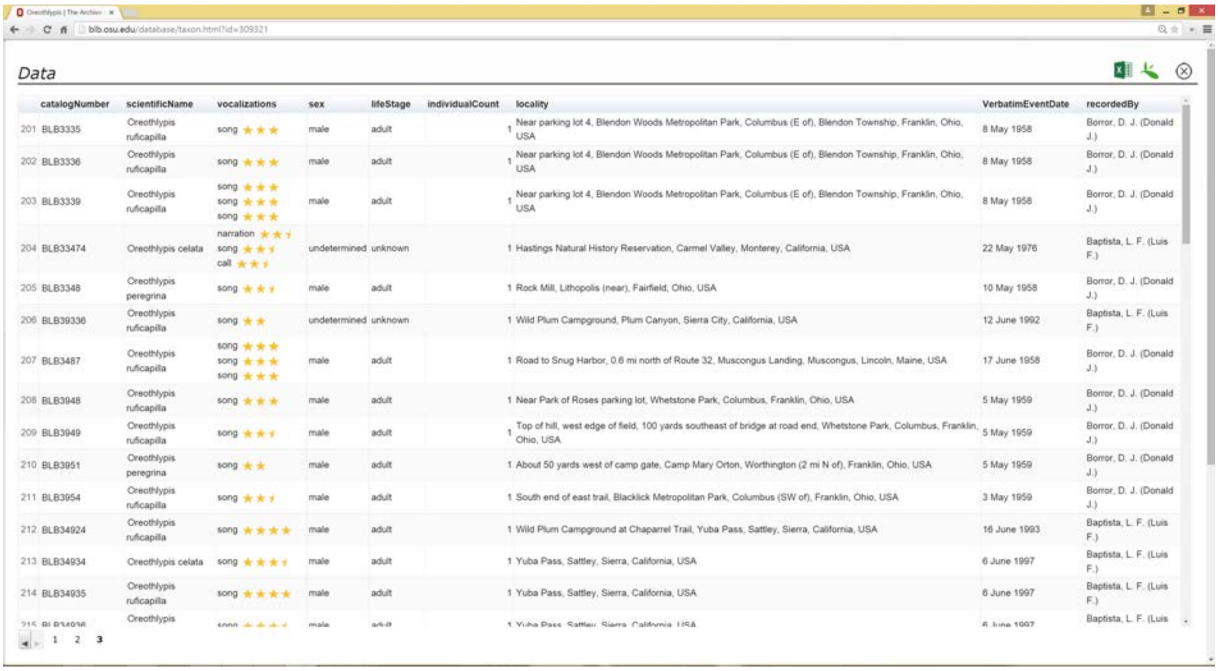
91 foreground recordings and **198** background recordings of *Poecile carolinensis* . Total recording duration 1:05:31.

Results format: detailed | concise | codes | sonograms

	Common name / Scientific	Length	Recordist	Date	Time	Country	Location	Elev. (m)	Type	Remarks	Actions	Cat.nr.
▶	Carolina Chickadee (Poecile carolinensis)	1:29	Andrew Spencer	2009-05-29	19:00	United States	near Rocksprings, Edwards Co., Texas	?	begging call	quiet begging calls from an older... more » [sono]	**A** B C D E	XC34950 🅭
▶	Carolina Chickadee (Poecile carolinensis)	0:55	Andrew Spencer	2009-04-27	13:00	United States	Prairie Lakes, White River NWR, Desha Co., Arkansas	?	call	very heavy wind noise partially... more » [sono]	**A** B C D E	XC33598 🅭
▶	Carolina Chickadee (Poecile carolinensis)	0:31	Chris Parrish	2008-03-11	18:21	United States	Sewanee, Franklin County, Tennessee	580	song	Songs: ss-see-t-dee [also] [sono]	**A** B C D E	XC18446 🅭
▶	Carolina Chickadee (Poecile carolinensis)	0:34	Chris Parrish	2008-03-11	18:18	United States	Sewanee, Franklin County, Tennessee	580	song	Songs: ss-see-t-dee [also] [sono]	**A** B C D E	XC18445 🅭
▶	Carolina Chickadee (Poecile carolinensis)	0:32	Chris Parrish	2006-10-13	18:00	United States	Roark's Cove, Franklin County, Tennessee	319	call	Calls, with and without higher... more » [sono]	**A** B C D E	XC15190 🅭
▶	Carolina Chickadee (Poecile carolinensis)	0:06	Chris Parrish	2007-05-12	19:08	United States	Sewanee, Franklin County, Tennessee	580	call	Burry calls: ps-tit de-de-de-de [sono]	**A** B C D E	XC15189 🅭
▶	Carolina Chickadee (Poecile carolinensis)	0:33	emganin	2015-04-21	06:00	United States	Grapevine, Tarrant County, Texas	170	song	before dawn at the pond behind the... more » Originally a mystery recording. See the forum. [also] [sono]	A **B** C D E	XC237815 🅭
▶	Carolina Chickadee (Poecile carolinensis)	0:08	Peter Boesman	1998-02-20	07:00	United States	Sequoyah NWR, OK		call	ID certainty 100%. (Archiv. tape 75... more » [sono]	A **B** C D E	XC227273 🅭
▶	Carolina Chickadee (Poecile carolinensis)	0:46	Peter Boesman	1998-02-20	07:00	United States	Sequoyah NWR, OK		call	ID certainty 100%. (Archiv. tape 75... more » [sono]	A **B** C D E	XC227272 🅭
▶	Carolina Chickadee (Poecile carolinensis)	0:32	Tom Auer	2015-01-01	07:33	United States	Washington, District of Columbia, District of Columbia	0	call, song	http://ebird.org/ebird/view/checklist?... more » [sono]	A **B** C D E	XC209081 🅭
▶	Carolina Chickadee (Poecile carolinensis)	0:20	Dan Lane	2014-05-02	08:30	United States	W Lancaster, Lancaster Co., Pennsylvania	90	call, song	Natural vocalizations from pair as they... more » [also] [sono]	A **B** C D E	XC179084 🅭

Figure 4.2: Example of information provided along with recording files in Xeno Canto [119].

The recordings are categorised based on the taxonomy name or geography location. They also provide some additional descriptions about each recording such as the quality level; sound types; and the location of the recordings (see figure 4.3). All the recordings in Borror's archive are classified into seven different recording quality levels, from poor to very good quality (see table 4.1).



Oreothlypis | The Archive
blb.osu.edu/database/taxon.html?id=309321

Data

	catalogNumber	scientificName	vocalizations	sex	lifeStage	individualCount	locality	VerbatimEventDate	recordedBy
201	BLB3335	Oreothlypis ruficapilla	song ★★★	male	adult	1	Near parking lot 4, Blendon Woods Metropolitan Park, Columbus (E of), Blendon Township, Franklin, Ohio, USA	8 May 1958	Borror, D. J. (Donald J.)
202	BLB3336	Oreothlypis ruficapilla	song ★★★	male	adult	1	Near parking lot 4, Blendon Woods Metropolitan Park, Columbus (E of), Blendon Township, Franklin, Ohio, USA	8 May 1958	Borror, D. J. (Donald J.)
203	BLB3339	Oreothlypis ruficapilla	song ★★★ song ★★★ song ★★★	male	adult	1	Near parking lot 4, Blendon Woods Metropolitan Park, Columbus (E of), Blendon Township, Franklin, Ohio, USA	8 May 1958	Borror, D. J. (Donald J.)
204	BLB33474	Oreothlypis celata	narration ★★⯪ song ★★⯪ call ★★⯪	undetermined	unknown	1	Hastings Natural History Reservation, Carmel Valley, Monterey, California, USA	22 May 1976	Baptista, L. F. (Luis F.)
205	BLB3348	Oreothlypis peregrina	song ★★⯪	male	adult	1	Rock Mill, Lithopolis (near), Fairfield, Ohio, USA	10 May 1958	Borror, D. J. (Donald J.)
206	BLB39336	Oreothlypis ruficapilla	song ★★	undetermined	unknown	1	Wild Plum Campground, Plum Canyon, Sierra City, California, USA	12 June 1992	Baptista, L. F. (Luis F.)
207	BLB3487	Oreothlypis ruficapilla	song ★★★ song ★★★ song ★★★	male	adult	1	Road to Snug Harbor, 0.6 mi north of Route 32, Muscongus Landing, Muscongus, Lincoln, Maine, USA	17 June 1958	Borror, D. J. (Donald J.)
208	BLB3948	Oreothlypis ruficapilla	song ★★★	male	adult	1	Near Park of Roses parking lot, Whetstone Park, Columbus, Franklin, Ohio, USA	5 May 1959	Borror, D. J. (Donald J.)
209	BLB3949	Oreothlypis ruficapilla	song ★★⯪	male	adult	1	Top of hill, west edge of field, 100 yards southeast of bridge at road end, Whetstone Park, Columbus, Franklin, Ohio, USA	5 May 1959	Borror, D. J. (Donald J.)
210	BLB3951	Oreothlypis peregrina	song ★★	male	adult	1	About 50 yards west of camp gate, Camp Mary Orton, Worthington (2 mi N of), Franklin, Ohio, USA	5 May 1959	Borror, D. J. (Donald J.)
211	BLB3954	Oreothlypis ruficapilla	song ★★⯪	male	adult	1	South end of east trail, Blacklick Metropolitan Park, Columbus (SW of), Franklin, Ohio, USA	3 May 1959	Borror, D. J. (Donald J.)
212	BLB34924	Oreothlypis ruficapilla	song ★★★★	male	adult	1	Wild Plum Campground at Chaparrel Trail, Yuba Pass, Sattley, Sierra, California, USA	16 June 1993	Baptista, L. F. (Luis F.)
213	BLB34934	Oreothlypis celata	song ★★★⯪	male	adult	1	Yuba Pass, Sattley, Sierra, California, USA	6 June 1997	Baptista, L. F. (Luis F.)
214	BLB34935	Oreothlypis ruficapilla	song ★★★★	male	adult	1	Yuba Pass, Sattley, Sierra, California, USA	6 June 1997	Baptista, L. F. (Luis F.)
215	BLB34936	Oreothlypis	song ★★★⯪	male	adult	1	Yuba Pass, Sattley, Sierra, California, USA	6 June 1997	Baptista, L. F. (Luis

1 2 3

Figure 4.3: Example of information provided along with recording files in Borror's archive [120].

Table 4.1: Seven different recording quality levels, from poor to very good quality, as classified in Borror's archive [120].

Score:	⯪☆☆☆☆	★☆☆☆☆	★⯪☆☆☆	★★☆☆☆	★★⯪☆☆	★★★☆	★★★★
Quality description:	Poor	Poor to fair	Fair	Fair to good	Good	Good to very good	Very good

4.3 Recent bird classification challenges with corresponding data sets

In recent years, a few international challenges [123, 124, 125, 126] have been taken place on the identification of birds based on their songs or calls. All these challenges proposed their tasks to be evaluated on their own collected and provided data set. Here is a summary of the four most notable bird classification challenges including a review of their corresponding data sets.

4.3.1 ICML4B 2013

As a joint contribution to 30th International Conference on Machine Learning in Atlanta in June 2013, the ICML4B bird challenge [123] provided researchers with a data set of 90 audio files and challenged them to develop an algorithm to identify 35 bird species among them. The algorithm was expected to be developed based on the up-to-date knowledge of machine learning.

Their provided data are available in [127] and it consists of a training and a testing data set. The training data includes 35 recording files with a total training duration of 18 minutes. The duration of each file is 30 seconds and it contains the song of one bird species after which the file is named. The test data are longer and larger than the training data and consist of 90 recording files with a total test duration of more than 3.5 hours. The data is provided by the Museum national d'Histoire naturelle [128]. The institute is well-known as one of the most respected bird survey institutions in the world. All the training and test data were recorded in 16bit wav format with a sampling frequency of 44.1 kHz.

4.3.2 MLSP 2013

At the 23rd MLSP workshop in August 2013, the 9th annual MLSP competition [124] was conducted as a bird classification challenge, comprising of 15 species and 79 participants. The data set used in this competition was comprehensively collected in real-life situations and field conditions. The participants were provided with a ten-sound audio recording and their task was to identify the bird species present in the recording by developing a classifier.

The new data set introduced in this competition comprises of 645 ten-second audio files in uncompressed mono WAV format, with a 16 kHz sampling frequency and bitrate of 16 bits per sample; it embodies 19 bird species. The sound files were collected over the years with several recording per day, taken usually around dawn when birds are more active.

The recordings have been inspected by a group of experts and have been labelled only with the name of the set of the species which they present. The inspections are performed by listening to the audio files and analysing their spectrograms; each expert provides a rating of their confidence in detecting the correct species along with each label set. Confidence weight majority voting is then used to form the final label set. The entire data including some relevant information about the competition are available in [129].

4.3.3 NIPS4B 2013

Taking place in Lake Tahoe, Nevada, in December 2013 during the Neural Information Processing Scaled for Bioacoustics (NIPS) international conference, the NIPS4B challenge [125] was the biggest bird classification challenge in 2013. The data provided by the Biotope Society [130], consists of two subsets of training and test data. The training data comprises 687 short recording files with the overall length of nearly 1.5 hours and it was annotated to 87 different sound classes of birds and their ecosystems. Apart from 53 different bird species, it includes 7 insect species and a batracian which were living alongside these birds. The testing data consists of 1000 sound files with the total length of nearly 2 hours and it contains all the species which exist in the training data. All the sound files are recorded as a mono WAV format file and sampled at 44.1 kHz with a variable duration from 1 up to 5.75 seconds. The entire data is available in [131] to download freely. The aim of the NIPS4B competition was to identify which of these 87 sound classes exist in the other 1000 test recording files. More than thirty teams participated in the NIPS4B 2013 challenge. A description of the entire competition along with the best provided systems are summarised in [125].

4.3.4 BirdCLEF 2014

BirdCLEF 2014 [126] was conducted with the partnership of the NIPS4B competition but improving and enlarging many of its aspects. The number of species employed in this competition was significantly higher and the data was recorded in real world situations with a large number of recordists. By using the metadata and defining information retrieval oriented metrics, they provide a more usage-driven and system-oriented benchmark. However, the huge diversity in the data collection conditions such as recording devices, recordists, context, background noise, etc. has raised the risk of confusion between different classes and as a result, the task of bird sound identification is notably harder. It will therefore probably produce substantially lower scores and offer a better progression margin towards building real-world generalist identification tools.

As mentioned previously, the entire data set available for this challenge needed to be divided into two sets of training and testing. To do so, one third of the observations of each species were randomly selected to move to the testing set and the rest were added to the training set. All the recordings of one species which had been recorded by one person on a single day were regarded as one observation, so had only be fitted into one of the two subsets.

The data sets used in BirdCLEF for both training and testing were all made up from audio files hosted on xeno-canto (XC) [119]. There are 501 species from Brazil as, with 14027 recordings, it has the highest number of recordings on XC. There is a range of 15 to 91 recordings per species, with 10 to 42 different recordists involved. The audio files are in wav mono format (16 bits), being normalised to the bandwidth of 44.1 kHz. Each audio file is associated with a set of metadata such as the type of sound (call, song, alarm, fight, etc.), the date and location of the recording, some common names and quality ratings.

In total, 87 research groups from all over the world attended this competition and their task was defined as capturing the most singing species in each file from the test set. The summary of their works and research is presented in [126].

4.4 Data description used in experimental evaluations

In this section the data sets that are used in all the experimental evaluations in this study are introduced. These data are provided by the Borror Laboratory of Bioacoustics' archive of bird sounds [120], where all the quality-based recording files are selected from the categories of 4 to 7 in Table 4.1.

As the diagram in Figure 4.4 illustrates, our data set consists of 50 different sub-sets of bird species (see appendix A) and each has been presented with several files. A total of 964 audio recording files are included in this data with an overall time of 38 hours, where each individual species has between 30 and 95 minutes of recording. This makes an average of 45 minutes of data per species, which is typically available in several recording files of between one and fifteen minutes each. All the sound files are recorded as mono

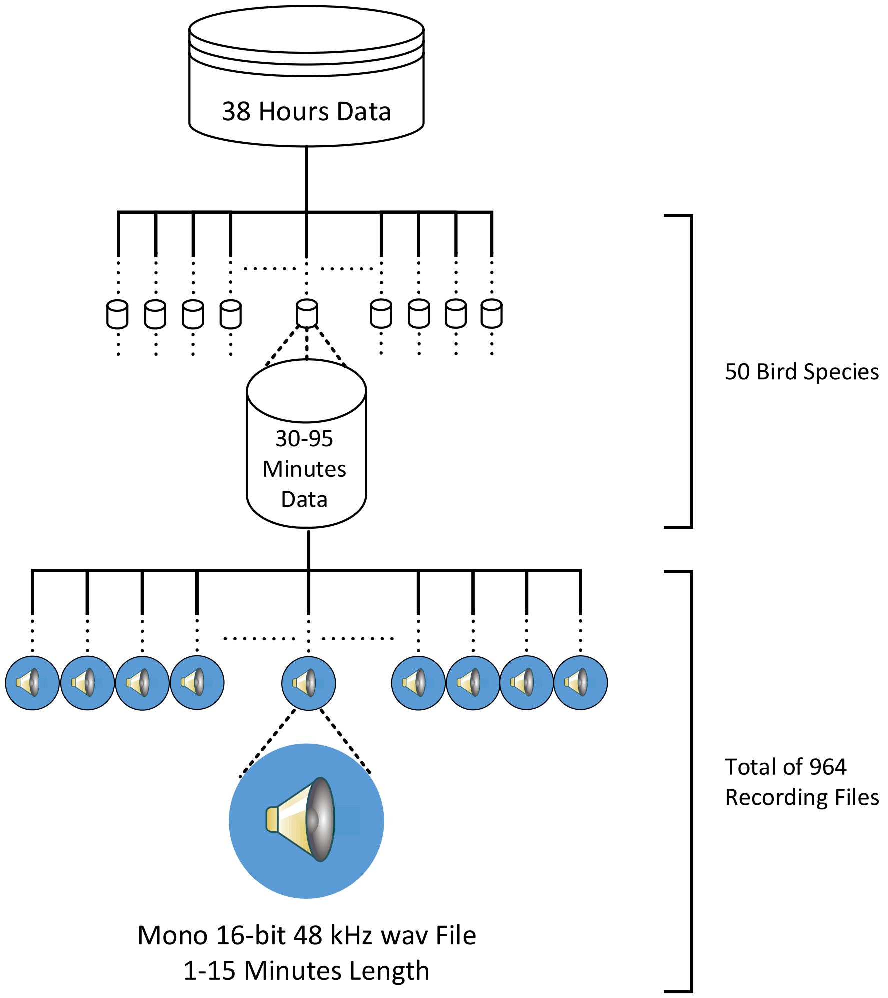


Figure 4.4: The collected data description provided by the Borror Laboratory

16-bit wav files with a sampling rate of 48 kHz and each recording is labelled with the name of particular bird that produced the main vocalisations . The information about each available bird species in our data, followed by the common name, total number of files and the total length of each species are presented in appendix A. Moreover, detailed information about each recording file including: the file ID number, the recording's length and the corresponding recording quality scores are presented in appendix B.

4.5 Manual annotation of recording files

As the data have been recorded in real world natural habitats of birds, along with the desired bird vocalisation sound, each recording may also contain some irrelevant information, including various background noises from other animals or humans. Each individual file is labelled with the name of its own bird species; however, there is no further annotation information available with the data to describe the content of the corresponding bird sound in the entire file. Furthermore, in order to evaluate the performance of the automatic segmentation and feature extraction procedures (see Chapter 5 section 5.4), and also to allow other researches to perform and evaluate comparative experiments with the real field corpus, this research aimed to annotate each individual recording of data manually into pre-defined audio sub classes.

To do this, first each individual recording file has been played and aurally inspected. During this manual inspection, each audible part of the signal is detected as a non-silent audio segment (or a sound event). Then, each detected audio segment is assigned to its relevant sound sub-class based on the nature of its content. Finally, all this annotation information with its corresponding time-stamps was written in a label text file.

4.5.1 Pre-defined labelling sound classes

As discussed above, in order to have a segment level annotation for each recording and based on the different possible sounds in the context, twelve labelling classes have been defined as follows:

1. *Bird vocalisations*: Since the name of each recording file is associated with a specific

bird species, all the vocalisation of that bird, including songs and calls, can be assigned into this class.

2. Other birds' vocalisations: Sometimes along with the sound of the main species that each individual file represents, there are audible sounds of other birds also in the file. In that case, all the parts related to the vocalisations of other species are labelled as bird vocalisation background noises.

3. Human speech: Many of recording files contain some human speech, as the recordist provided some useful information about the study/observation in speech on the recording file. Most of the speech testimonies are about the species name, behaviour of the subject, date and time of the recording and the location, habitat description and the distance to the species. All the discovered speech segments are labelled as the human speech class.

Figure 4.5 illustrates an example of a human speech segment along with the bird vocalisation segment within a Borror's sound file.

4. Flying insects' sounds: This annotation class consists of sounds which are produced by flying insects and bugs. As the sound files are recorded in nature, data may contain the buzzing-like sounds which are produced by flying insects including bees and flies.

5. Other animals' sounds: It is typical to hear sounds of other animals near a bird's habitat in nature. There are plenty of sound files which contain sounds of other animals, such as frogs, cows, pigs, and sheep; these all are labelled as in this category.

6. Microphone and recording noises: Since all the data is recorded by various types of recording devices or microphones, the audio signals may be affected by the different types of microphone or audio noises. All these affected parts are annotated in this class. As it

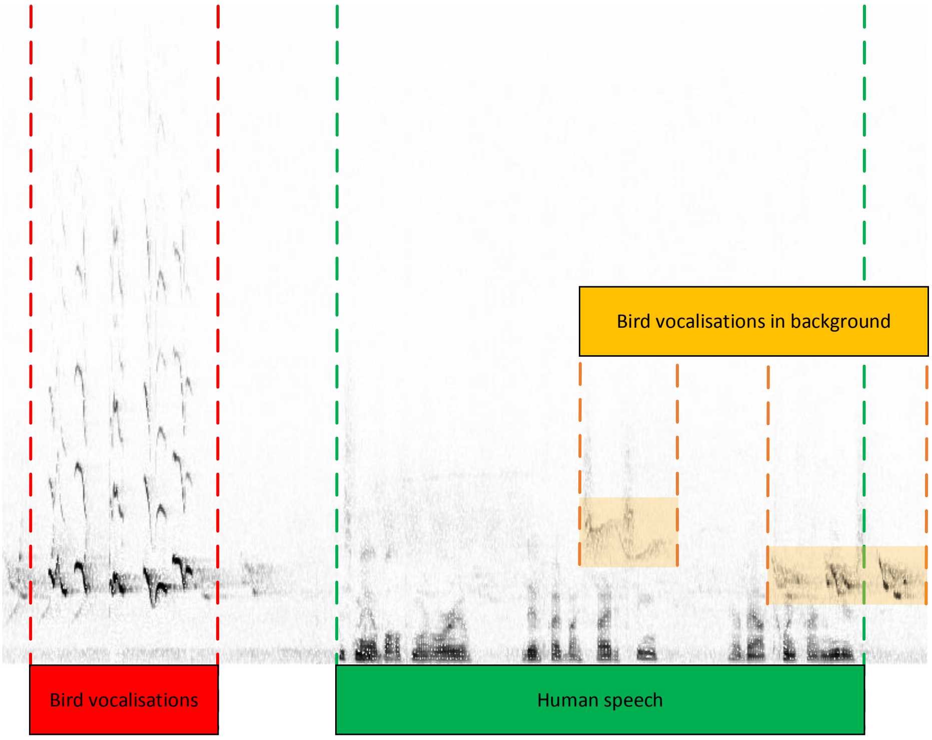


Figure 4.5: Sample of a speech segment in a Borror's recording file; recordist saying: "This male Baltimore Oriole, the side by side as seen, and about 25 meters ..." .

can be seen in Figure 4.6, a part of the Borror's recording is affected by a noise which is created by an ordinary portable recorder. Hence, the first and third segments are labelled as bird vocalisation and the second segment is labelled as microphone noise.

7. Tap sounds: All tap-like sounds such as hard knocking, tapping a hammer or dropping sounds are examples of sounds that can be labelled as in this category.

8. Wind sound noises: Many of the sound files were recorded in windy fields and consequently their recording signal contains some high wind background noise. Therefore, any part of the signal which is affected by high wind background noise is labelled as in this class.

9. Water sound noises: All the water sound noises include waterfalls and pouring rain

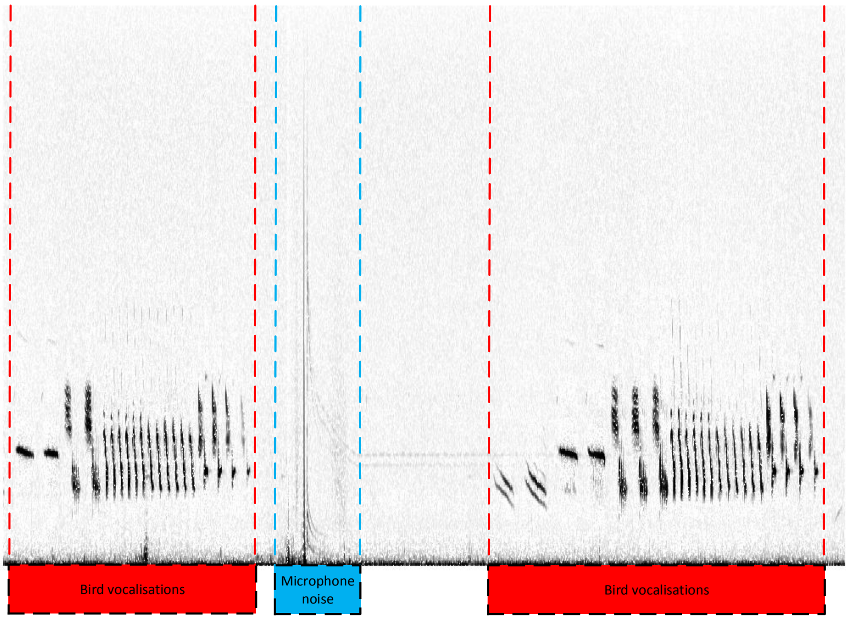


Figure 4.6: Sample of a microphone noise segment in a Borror's recording file

sounds and are recognised as water sounds in our data.

10. Motors and road noise: In some recording files, the bird's nest was located close to the road and the recording files are affected by several road background noises including car engines. Also, in some recordings there are a few low flight background noises audible. So any road or flight background noises are categorised as the motoring sound class.

11. Footsteps' sound noise: One of the main undesirable sounds which affect most of the recordings is the inevitable noises created by the recordist him/herself. During the recording process, recordists might need to change their position by walking or moving towards their subject. Hence, all the walking or moving background noises are labelled as in this class.

12. Other background noises: There are other types of background noises existing in the data, such as clock bell sounds, alarm sounds, siren and chainsaw sounds. As they happened rarely in the data, all the other minor background noises have being labelled as in this class.

For the sake of simplicity and speed in the labelling process, a unique keyword is assigned to each class which is used in annotation process (see table 4.2).

Table 4.2: The keywords associated with each of the 12 predefined annotation classes.

	Name of class label	Keyword name
1	Bird vocalisations	BIRD
2	Other birds vocalisation noises	BIRDBACK
3	Human speech	SPEECH
4	Flying insects’ sounds	FLYINS
5	Other animals’ sounds	ANIM
6	Microphone and recording noises	MICREC
7	Tap sounds	TAP
8	Wind sound noises	WIND
9	Water sound noises	WATER
10	Motors and road noise	MOTOR
11	Footsteps’ sound noise	WALK
12	Other background noises	NOISEOTH

4.5.2 Single and multiple labels

To perform the process of labelling, first all the existing segments of each recording signal need to be detected and marked. Then each segment will be labelled according to its comprising components. Two major situations may be encountered while inspecting each segment: there might be parts of the signal which represent only one single class of sounds; in that case, whether it is the relevant bird sound or a sound from any of the other pre-defined classes, the corresponding part will receive one single label. This procedure is called “single labelling”in our study.

On the other hand, there may be parts of the segment where two or more sounds from different pre-defined classes overlap each other. Consequently, that segment will receive several parallel labels based on its content. This situation is called “multiple labelling”. Figure 4.7 demonstrates a possible situation of multiple labelling; the current segment is divided into three parts, two of which have received one single label of “Other bird vocalisation in background”; while the middle part of the segment is labelled as a combination of “Bird vocalisation”and “Other bird vocalisations in background”.

4.5.3 Labelling the data using Transcriber software

There are many free licence annotation tools available publicly on the Internet for assisting with the manual annotation of audio and speech signals. However, a proper software environment is essential for annotating a vast number of recorded data. Transcriber is a user-friendly graphical annotation tool which can be used for segmenting, labelling and transcribing recorded speech [132]. It supports most standard audio file formats including wav files and the annotation output can be provided in text file format. This software has

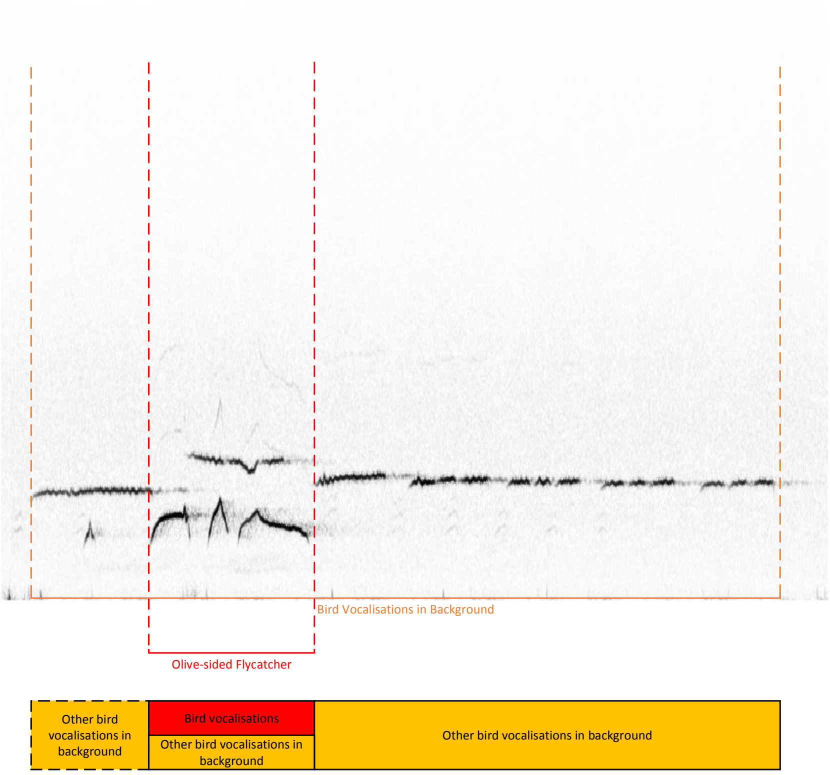

Bird Vocalisations in Background
Olive-sided Flycatcher
Other bird vocalisations in background
Bird vocalisations
Other bird vocalisations in background
Other bird vocalisations in background

Figure 4.7: Example of a multiple labelling task in the annotation procedure

a simple one-step installation and is available for Windows, Linux and Mac OS operating systems.

In this work, Transcriber version 1.5.1 has been used for annotating our selection of the Borror data set.

1. Pre-setting and importing the recording files

The manual bird sound annotation process for each individual file starts with importing the recording file into the Transcriber software. This can be done by dragging the file into the Transcriber's interface or importing it through the menu bar. Subsequently, the recording signal can be viewed in the signal window (see figure 4.8) with pre-set customised

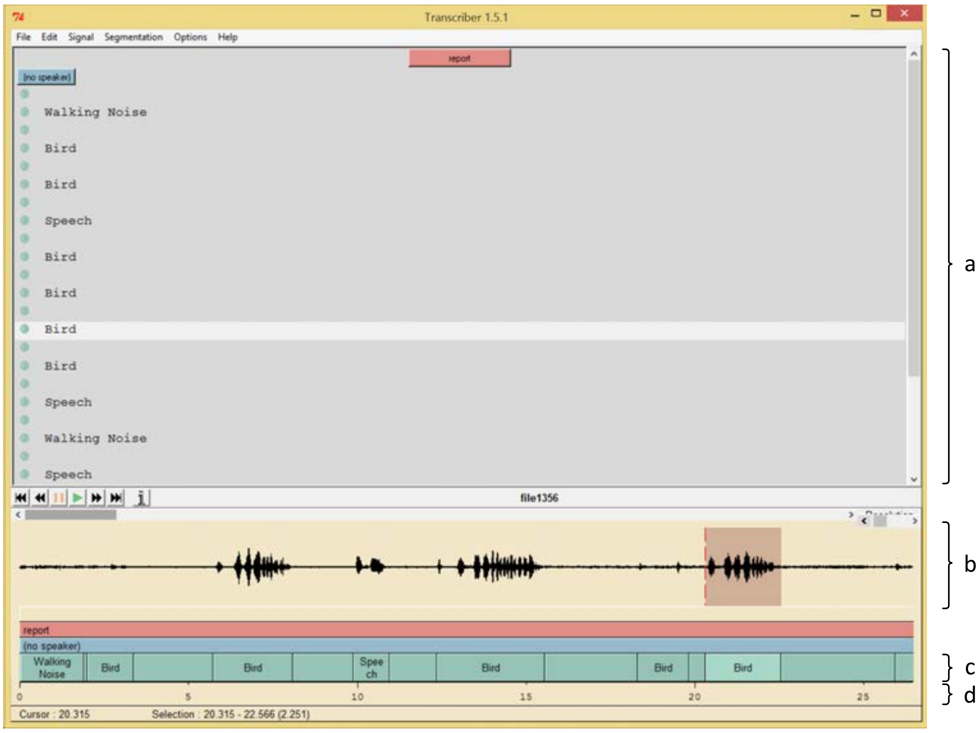


Figure 4.8: Transcriber interface: a) Transcription window: It is synchronised with the segmentation panel and the label name of each segment can be typed in a corresponding text editor line. b) Signal view panel with red curser: It shows the input signal and it is synchronised with the time bar and segmentation panel. Display resolution can be changed from 1 second to a maximum 5 minutes. The red curser shows the current position on the signal. c) Segmentation panel: It shows all the segments' blocks and their label names which are synchronised with the transcription window. d) Time bar.

time resolution. The next steps are as follows:

2. Segmenting the signal

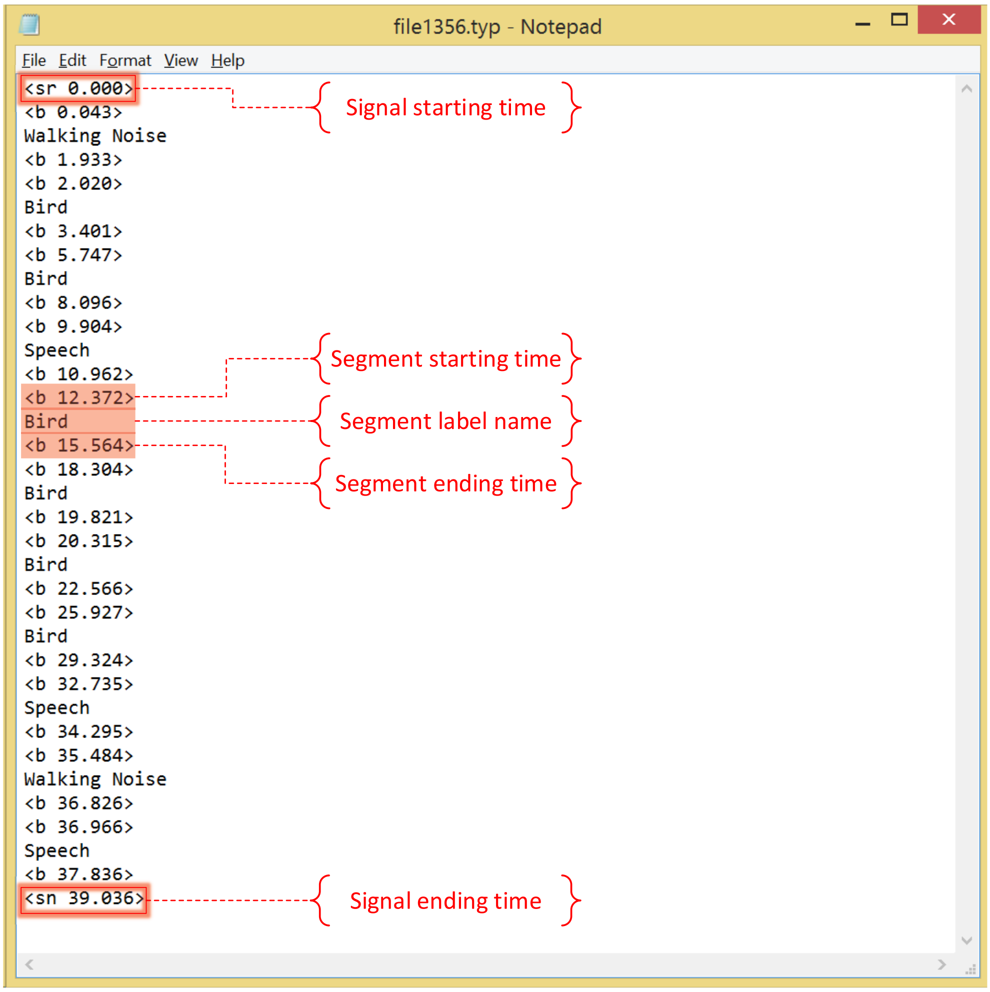
By listening to the recording signal, the audible parts (non-silent parts) of the signal can be extracted as sound events. Then, each sound event is marked as a new segment by defining its boundaries on the timeline in the signal's window (see figure 4.8).

3. Labelling the segments

Each detected segment will then be labelled with one or more of the pre-defined subclasses. This can be done by typing the corresponding label title of each segment in the transcription text editor section (see figure 4.8). In the case of multiple labelling, in which each segment needs to be labelled with more than one class, all the relevant parallel sound classes will be mentioned, separated by the character '/' in the transcription window. The silence parts of the signal should remain empty in the transcription window. Adding, splitting, or removing any part of the signal can be easily carried out by a simple mouse manipulation in the signal view panel. With just a single keystroke, the user can easily select, zoom and play any portion of the signal; and also pause and restart in the case of playback.

4. Exporting the annotation file

At the end of this process, all the annotation and labelled objects with their transcription information are exported into a single textual output file, as .typ text format. As it can be seen in Figure 4.9, the first piece of information provided in each output file is the starting time of the signal as header information e.g. <sr 0.000> and the example finishes with pointing out the signal's end time, e.g. <sn 39.036>. Then, each detected labelled segment is described by showing the starting and ending time of the segment along with its corresponding annotation name.



```
<sr 0.000>
<b 0.043>
Walking Noise
<b 1.933>
<b 2.020>
Bird
<b 3.401>
<b 5.747>
Bird
<b 8.096>
<b 9.904>
Speech
<b 10.962>
<b 12.372>
Bird
<b 15.564>
<b 18.304>
Bird
<b 19.821>
<b 20.315>
Bird
<b 22.566>
<b 25.927>
Bird
<b 29.324>
<b 32.735>
Speech
<b 34.295>
<b 35.484>
Walking Noise
<b 36.826>
<b 36.966>
Speech
<b 37.836>
<sn 39.036>
```

Figure 4.9: Example of Transcriber output text file (.typ file)

4.5.4 Scoring the quality of the recordings

As mentioned previously in the data description section, all the recording files used in this data set are selected from the quality categories of 4 to 7 of Table 4.1. Apart from the above-mentioned annotation process and instead of using the Borror's own quality rating, two different rating scores have been incorporated; namely, the recording quality score and background noise level score, in order to express the overall quality of each recording in a much more precise way. All the obtained scores are available in appendix B, along with the file ID and the length of each sound file.

1. Recording quality score

As the data was recorded by different recordists and with different recording devices, a recording quality score to describe the quality of the recording of each sound file was devised. This score is a five-point scale rating score and is obtained manually by listening to the entire signal. Table 4.3 shows all the quality levels with their corresponding score value.

Table 4.3: Quality of recording rating scale

Score:	1	2	3	4	5
Quality description:	Poor	Poor to fair	Fair	Good	Very good

2. Background noise level score

Due to the existing background noises along with the recordings, the background noise level score is the measurement rating that expresses the overall background noise in the entire recording files. It is a six-point rating scale which demonstrated in Table 4.4. The score is rated as zero if there is no background noise or other sound classes (except of bird vocalisation) available in the recording file. Otherwise, it is rated from 1 to 5 with regards to the overall amount of multiple background noises.

Table 4.4: Background noise rating scale

Score:	0	1	2	3	4	5
Background noise level description :	No background noise	Very low background noise	Low background noise	Medium background noise	High background noise	Very high background noise

4.5.5 User interface

As mentioned previously, the Transcriber output file includes some useful information, such as the annotation name of each detected segment with corresponding starting and ending points. There is also some more relevant information available along with Transcriber's annotation information (see Figure 4.10), such as the corresponding bird species name and the file ID of each sound file, extracted directly from the Borror data information. Moreover, the recording quality score and background sound level score, which have been obtained manually during the annotation process, are also included in each labelling file.

As mentioned earlier, the Transcriber output file includes some useful information such as the annotation name of each detected segment with corresponding starting and ending points. There are few more relevant information available along with Transcriber annotation information (see Figure 4.10), such as the corresponding bird species name and the file ID of each sound file, extracted directly from the Borror data information. Moreover, the recording quality score and background sound level score which have been obtained manually during the annotation process is also included in each labelling file.

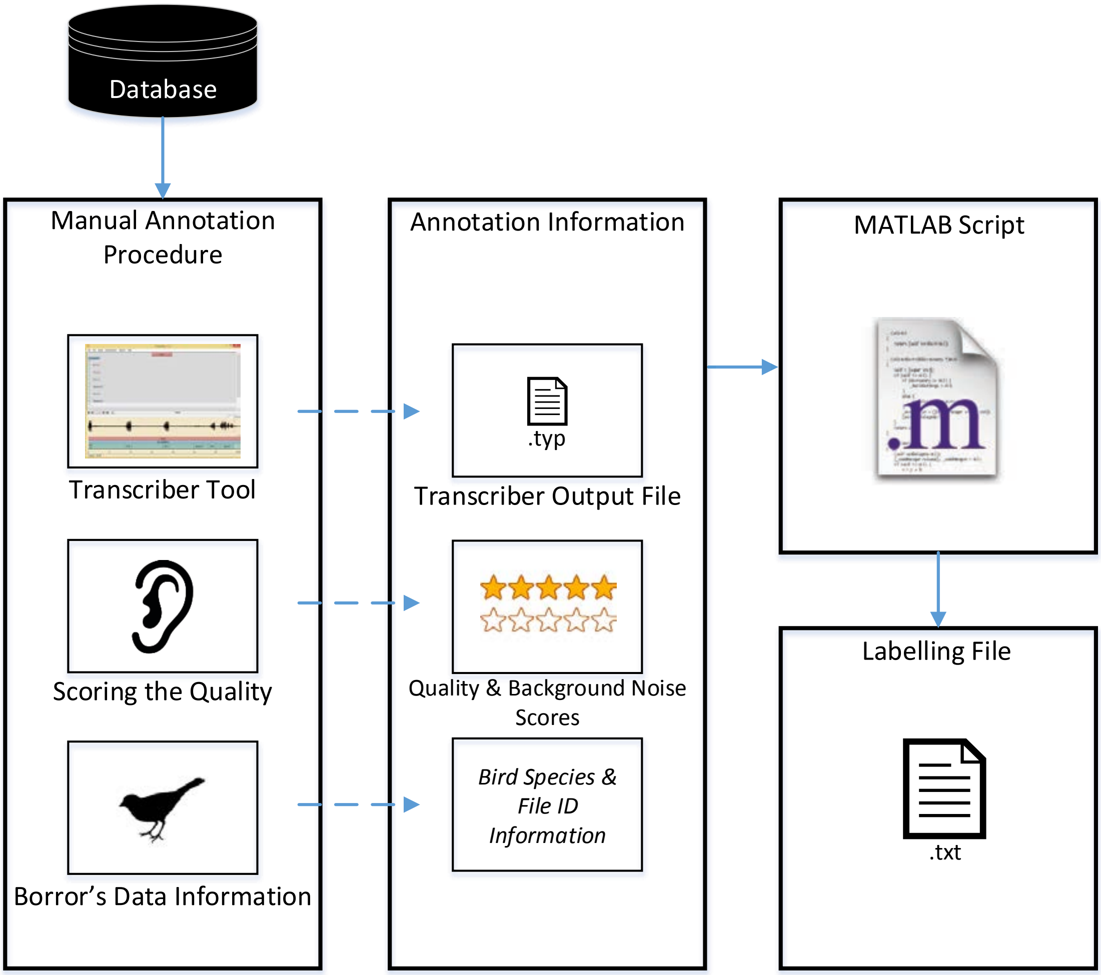


Figure 4.10: Summary of the annotation process

For the purpose of creating a text grid labelling output file for each individual recording file in our data, which accommodates all the above annotation information, and in order to refine and customize the output file, a MATLAB script has been developed.

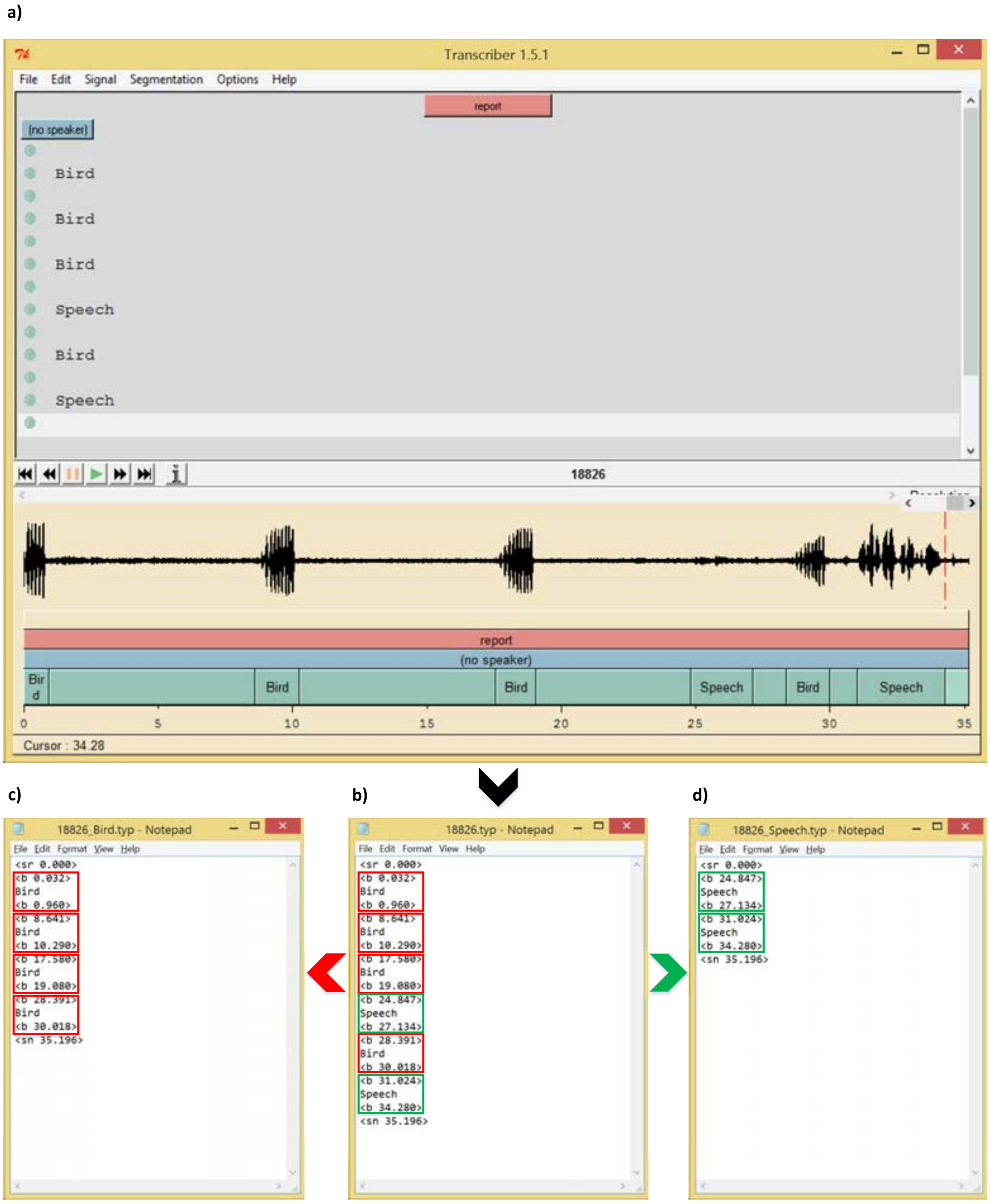
The MATLAB script provides several flexibilities to the users in terms of customising the labelling files as follows:

1. Modifying the name of each pre-defined annotation class

The script provides the facility for users to modify the name of each annotation class based on their preferences. For example, instead of ‘bird vocalisations’ they may use their own words such as ‘Bird’ or ‘B’.

2. Exporting multiple label files

One of the main features of this MATLAB script is that it provides users with the facility of managing their annotation classes based on their requirements and preferences. This includes the possibility of having individual label files for each detected annotation class within a sound file, rather than the default option of having one single label file per sound file. For instance, as it is demonstrated in Figure 4.11, the example recording file from the data set with the approximate length of 35 seconds has been labelled in six different segments by Transcriber, in which the first three and the fifth segments (red coloured) have been labelled as ‘bird’ and the fourth and sixth segments (green coloured) have been labelled as ‘speech’. The user has the choice of storing the information of all the detected segments in two separate label files named ‘18826_Bird.typ’ and ‘18826_Speech.typ’, each containing the relevant information of their corresponding annotation class. It even provides users with the facility of only including their preferred annotation classes in the output file and eliminating others. It only requires the entering of the desired classes in the function’s setting prior to the processing.



c) 18826_Bird.typ

```
<sr 0.000>
<b 0.032>
Bird
<b 0.960>
<b 8.641>
Bird
<b 10.290>
<b 17.580>
Bird
<b 19.080>
<b 28.391>
Bird
<b 30.018>
<sn 35.196>
```

b) 18826.typ

```
<sr 0.000>
<b 0.032>
Bird
<b 0.960>
<b 8.641>
Bird
<b 10.290>
<b 17.580>
Bird
<b 19.080>
<b 24.847>
Speech
<b 27.134>
<b 28.391>
Bird
<b 30.018>
<b 31.024>
Speech
<b 34.280>
<sn 35.196>
```

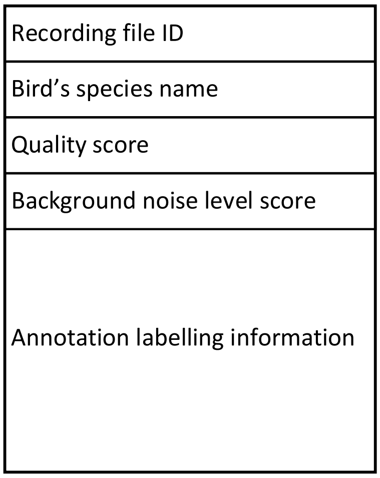
d) 18826_Speech.typ

```
<sr 0.000>
<b 24.847>
Speech
<b 27.134>
<b 31.024>
Speech
<b 34.280>
<sn 35.196>
```

Figure 4.11: a) Transcriber interface, b) Transcriber's output file, c) Bird label file for sample file 18826, d) Speech label file for sample file 18826.

3. Modifying the output structure

The user can also modify the structure and format of the output file to include their own preferred information. By default, the output file has the structure shown in Figure 4.12 including this information: file ID number, bird's species name, quality score, background noise level score and annotation labelling information and it is exported as a '.txt' file.



Recording file ID

Bird's species name

Quality score

Background noise level score

Annotation labelling information

Figure 4.12: Default structure of the output file

4.6 Conclusion

In this chapter, first the available large data files of bird sounds were introduced as well as the current demand for their associated labelling and annotation information, which leads to the work conducted in this part of the research. Then the available data set used in recent bird classification challenges was briefly discussed. Also, the specific data set used in this thesis was introduced. The first step of the annotation was performed manually by listening to each recorded sound file. In the next step, each file was segmented and labelled in its comprising annotation classes by the Transcriber software. Moreover, each sound file has been investigated manually regarding its recording quality and the background noise level and has been rated based on this information. Lastly, a user interface has been designed as a MATLAB script which enables the user to manage the label files based on their preferences. In order to improve the annotation process and as a future plan to develop this research further, more specific annotation classes can be defined and employed in labelling to increase accuracy.

Chapter 5

SEGMENTATION AND ESTIMATION OF ACOUSTIC FEATURES FOR BIRD VOCALISATION

5.1 Introduction

As mentioned previously in Chapter 2, the first step of a bird sound identification system is to split the continuous acoustic signal into smaller isolated units (segments), where sound events are present, along with extracting relevant information from these segments into a compact set of features which can represent the characteristics of the bird species. In this chapter, with the assumption that the bird's vocalisation is tonal and for the purpose of feature extraction and the signal's segmentation, the algorithm introduced and presented in [5, 6] has been employed with some modifications that we introduced in [133, 134].

The authors in [5, 6] introduced a method for the detection of sinusoidal signals of speech and sound signals in a noisy environment. In this thesis, their proposed sinusoidal detection algorithm was employed in order to obtain the frequency tracks' segments for bird vocalisation parts in the data. Further modification and refinement steps were also presented (see section 5.3.3) to deal with the background noises existing in the data (described in Chapter 4).

Sinusoidal components are fundamental building blocks of sound. Based on Fourier's theorem, every sound can be presented as a sum of sinusoids having fixed amplitude and frequency, but this is a highly inefficient model for non-tonal and changing sounds. In other words, any tonal sound can be modelled efficiently as sum of windowed sinusoids over short stationary time, and a spectral peak is naturally modeled as a sinusoidal component which has been shaped by some kind of window function or amplitude envelope in the time domain [135].

In this study, the segmentation of the audio signal is performed based on detecting sinusoidal components in the signal. The detection of sinusoidal component is taken as a pattern recognition problem. It is performed on a signal frame basis. Since spectral peaks in the short-time spectrum are effective in capturing the tonal aspects of a sound signal, each peak in the spectrum of a signal frame is considered as a potential sinusoidal component. A set of features, extracted from the spectrum, is then obtained for each spectral peak. In the recognition stage, the detection's decision is based on the comparison of the distribution of the feature vectors of these potential sinusoidal components with our large collection of features corresponds to spectral peaks of noise and sinusoidal signals. At the end of the detection, a segment would be a sequence of these detected sinusoids. Furthermore, each obtained segment is characterised as a sequence of frequencies of these sinusoids at each frame signal. This continuous sequence is named as a frequency track segment, for the purpose of simplicity in this thesis.

In the following section, the procedure for the estimation of frequency tracks will be discussed in more detail, along with the further refinement of the frequency track extraction results.

5.2 Estimation of frequency tracks

5.2.1 Obtaining spectral magnitude shape and the phase continuity

As the short-time spectrum can be expressed by a set of features such as, spectral magnitude and the phase continuity information, the purposed method in [5, 6] uses both types of spectral information to obtain each spectral peak.

$S_l(k)$ denotes the short-time spectrum of the l^{th} frame of the signal and the k_p denotes the frequency index of a detected spectral peak in the short-time magnitude spectrum. A multivariate feature vector y is extracted for each peak representing both the spectral magnitude shape and phase continuity information around the peak [5, 6]. To obtain the magnitude shape feature, M number of magnitude spectral's point is used over the range of frequency bins from $k_p - M$ to $k_p + M$ as follows [134]:

$$y^1 = (|\tilde{S}_l(k_P - M)|, ..., |\tilde{S}_l(k_P - 1)|, |\tilde{S}_l(k_P + 1)|, ...|\tilde{S}_l(k_P + M)|)$$

Where $|S_l\tilde{(}k)|$ is the normalised spectral magnitude and it obtained by $|S_l(k)|/|S_l(k_p)|$; where $|S_l(k_p)|$ is the magnitude value at the peak and M is the number of bins considered around the peak.

The phase continuity features are obtained by using the spectral phase difference values over the range of frequency bins from $k_p - M$ to $k_p + M$, as follows:

$$y^2 = (\Delta\phi_l(k_p - M), \ldots, \Delta\phi_l(k_p - 1), \Delta\phi_l(k_p + 1), \ldots, \Delta\phi_l(k_p + M))$$

[134]

Where the phase difference between the current and previous signal frame is defined as $\Delta\phi_l = \phi_l(k) - \phi_{l-1}(k) - 2\pi kL/N$, as $\phi_l(k)$ and $\phi_{l-1}(k)$ are the phase of the frequency point k at frame-time l and $l-1$, respectively; and the term $2\pi kL/N$ is included in order to compensate for the shift of the sinusoidal signal between the adjacent signal frames, with L being the frame-shift in samples [6, 53].

5.2.2 Probabilistic modelling

The corresponding magnitude shape and phase continuity information of each detected spectral peak can be represented by the distribution of the multivariate feature vector $y = (y^1, y^2)$. In this section GMM is used to model this distribution.

The training process provided a large collection of features of y, corresponding to spectral peaks of noise and sinusoidal signals at various SNRs [6]. These data will be used to estimate the GMM parameters of noise, λ_n and of sinusoidal signal, λ_s. As mentioned previously, the maximum likelihood criterion is then employed to decide whether a spectral peak corresponds to signal or noise. For instance, if $p(y|\lambda_s) > p(y|\lambda_n)$, the peak will be detected as sinusoid [6].

The above provides a set of detected sinusoidal components at each signal frame; the signal-frames in which no sinusoid is detected indicate no presence of tonal bird vocalisation. Each continuous sequence of these detected frames is considered as an isolated segment. Then, each frequency track segment is presented as a sequence of frequencies of that sinusoid at each frame signal.

5.3 Experimental evaluation

5.3.1 Experimental procedure

The data used in this approach is the exact data that was introduced in Chapter 4, with the initial Borror's labelling files. This means that each existing recording file is labelled only with the name of the main bird species and the further obtained annotation labelling files (annotated data, as described in Chapter 4 section 4.5) are used to evaluate the quality performance of the entire automatic detection procedure (see section 5.4).

As can be seen in Figure 5.1, in this chapter all the available recording files within each subset data (all the data within the 50 bird species) are being used as a means of feature extraction. Therefore, at the end of this experimental evaluation, a set of frequency tracks sequences will be obtained for each recording file and these detected sequences will be used in further evaluations as the temporal data.

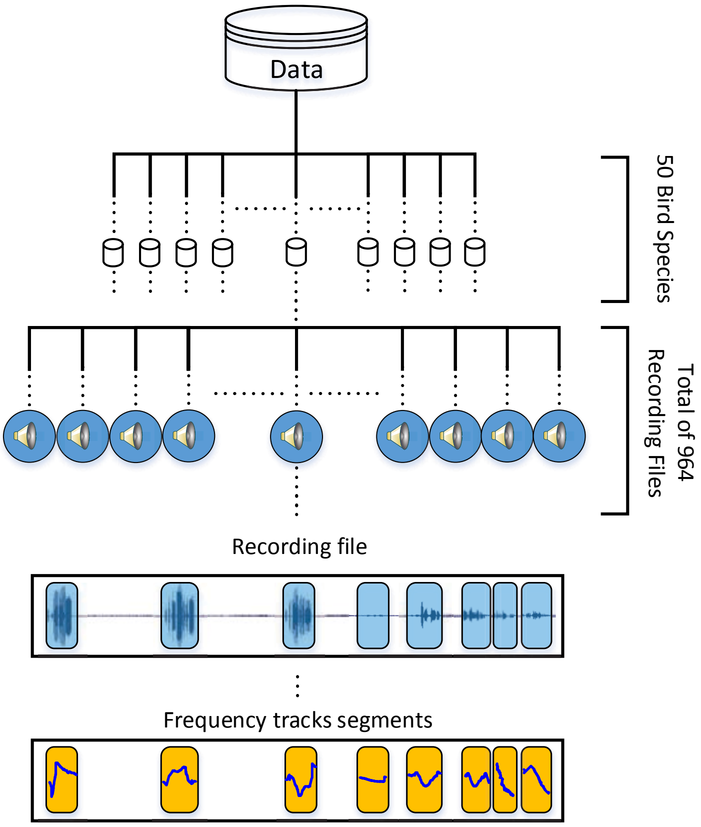


Figure 5.1: Data sets used in experimental evaluation procedure

5.3.2 Parametric setup

As the entire data was recorded with the sampling rate of 48 kHz, each acoustic signal is divided into frames of 256 samples corresponding to 5.3ms. There is also a shift of 48 samples between the adjacent frames which corresponds to 1ms. In accordance with the outcomes of some previous research [3, 133], the frame length is chosen to be shorter than that which is common in most bird processing studies. The use of longer frame lengths would provide better frequency resolution, however, due to the fast frequency variations in bird vocalisations, this would lead to some smearing in the spectrum. Then, a rectangular analysis window is applied to the signal and in order to provide a finer sampled Discrete Fourier Transform (DFT) spectrum, a $512 - point$ DFT is used, which means the signal is added by 256 zeros. The parameter M (section 5.2.1) is set to 6 frequency bins. The training of the models of the sinusoidal signals was performed using simulated sinusoids, with a range of linear frequency modulation. The models consist of 32 Gaussian mixture components.

5.3.3 Refinement of the detection results

The outcome of the sinusoidal detection method can be regarded as an initial segmentation of the acoustic scene. The following steps are presented, in a conservative way, to further refine this detection result, in which the frequency tracks' segments are obtained only for the parts in the signal where bird vocalisation exists and excluding all other background sounds and speech signals.

1-Discarding very short segments with a length less than 4 frames:

All the detected segments with a considerably short length, in this case less than 4 frames, will be assumed to be detected by errors. As such, all these very short segments are discarded in the first step.

2-Interpolating the segments:

For all the segments that are apart from each other for up to two frames and two frequency bins, an interpolation is performed between the beginning and the end point of the segments. This will avoid the accidental split of a segment due to a missed detection of a few frequency bins.

3-Discarding all the short segments with a length less than 14 frames:

Next, all the segments with a length of less than 14 frames are discarded. These short-length segments are unlikely to include any bird vocalisations in our data.

4-Excluding background co-vocalisations' segments:

As mentioned previously, the data which has been used here are recorded from natural bird habitats and they include some co-vocalisation of other birds or animals with the recordings. As this research is not interested in employing and modelling these background co-vocalisations' segments, the assumption has been made that the birds' species vocalisations are of a higher energy compared to any other background vocalisations in the recording. As a result, all the segments where their average energy were 15 dB below the highest average segment energy in each recording have been eliminated.

5-Excluding the human speech segments:

Finally, as the bird vocalisation segments correspond to the frequency regions greater than 2 kHz in this data, and in order to get rid of human speech vocalisations presented in the recordings, all the segments with a median frequency less than $2kHz$ have also been discarded.

Figure 5.2 shows a sample spectrogram of an audio field recording from the Borror data containing concurrent vocalisations of two bird species and the estimated frequency tracks before and after applying the refinement procedure. As can be seen, the detected frequency tracks (before the segmentation) clearly fit into corresponding birds vocalisations including even the weakest; examples of which are high frequency components around frequency index 120 and around frame time index 560, 600 and 1050. As listening to the recordings also reveals, these were related to the co-vocalisations of other birds' in the background. Furthermore, the desired birds vocalisations were all captured well in the final frequency.

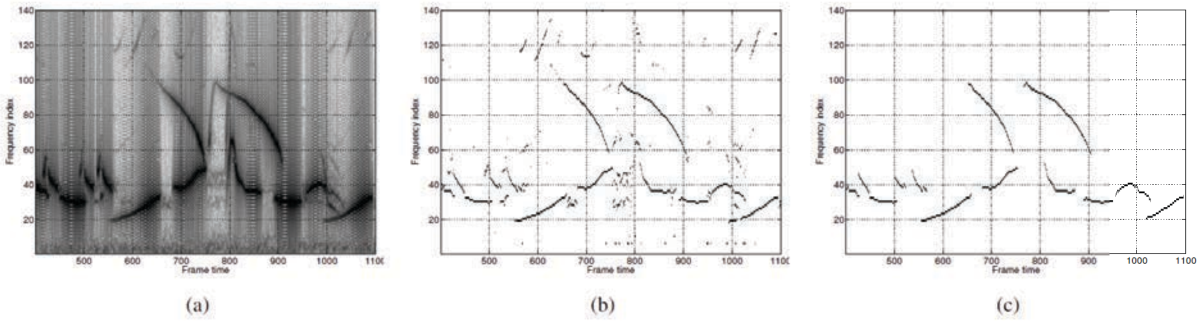


Figure 5.2: A sample spectrogram: (a) of audio field recording and the corresponding estimated frequency tracks, initial (b) and final (c).

5.4 Quantitative evaluations of the performance of the frequency tracks' detection system

This section provides two quantitative analysis methods to evaluate the quality performance of the entire automatic feature detection system, in terms of feature tracking and segmentation purposes.

5.4.1 Database description

In order to evaluate the quality of the entire automatic detection system, it is necessary to have a reliable reference for further analysis steps. All the quantitative evaluations were performed using some of the ground truth annotated data from Chapter 4 (as reference labels) to compare with the final outcome of the frequency tracks' estimation system.

The data set that is used in this section consists of six different subsets of bird species which are randomly selected from the available 50 sets of bird species. As Table 5.1 shows, a total of 88 variable-length audio recording files are included in this selected data set with an overall time of 4.5 hours, where each individual species has between 32 and 95 minutes of recording. Typically, this (selected) data set is about 14% of the entire data which is presented in Chapter 4.

Table 5.1: Datasets used for quantitative evaluations of the performance of the frequency tracks' detection system

Name of bird species	Number of recording file	Data length (minutes)
Bird 1: Carolina Wren	19	95
Bird 9: Louisiana Waterthrush	15	36
Bird 20: Hooded Warbler	11	41
Bird 31: Prothonotary Warbler	12	32
Bird 35: Kentucky Warbler	19	34
Bird 43: Savannah Sparrow	12	34
Total:	88	272 (4.5 hours)

5.4.2 Obtaining the detected output signal

As it is mentioned in section 5.1, we consider that birds produce tonal vocalisations. Furthermore, the estimation of the frequency tracks method provides a set of detected sinusoidal components at each signal frame; the signal-frames in which no sinusoid is detected indicate no presence of tonal bird vocalisation. In other words, the procedure for obtaining the frequency track performs similarly to a two-layered classifier, which labels the entire signal as two separate classes: bird vocalisation and non-bird vocalisation (includes silences). Hence, as can be seen in Figure 5.3, the detection outcome in each recording file is transformed to a binary transcription signal; where the one's value indicates bird vocalisation and the zero's value indicates non-bird vocalisation on each frame of the signal. For the sake of simplicity, this binary transcription signal is named "detected output signal" in this chapter.

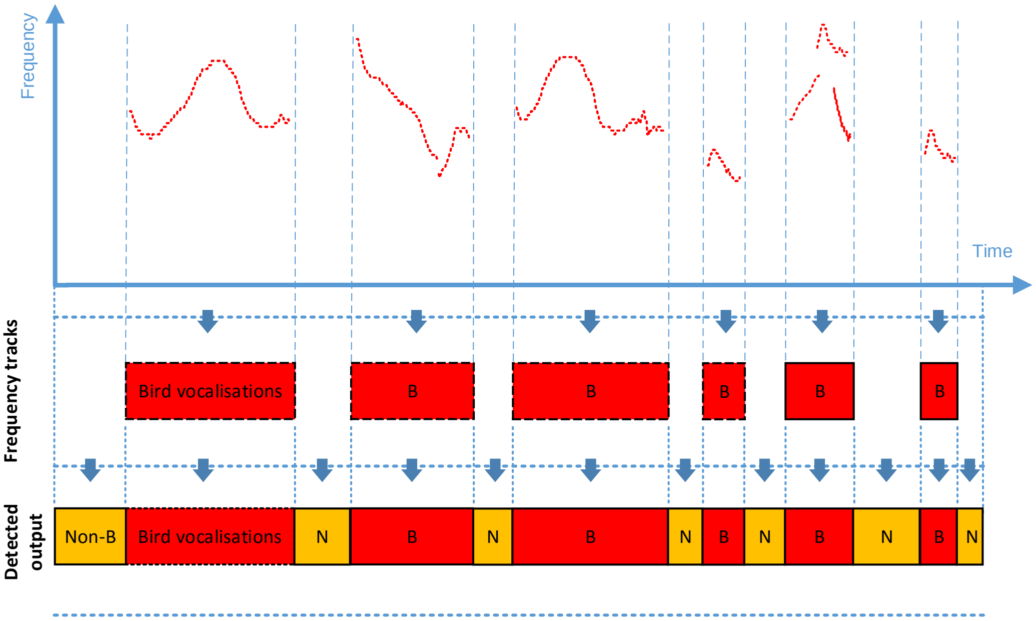


Figure 5.3: Obtaining the detected output signal, for each recording file, from the outcome of the frequency tracks' estimation procedure

5.4.3 Obtaining the reference transcription signal

According to the manual annotation procedure (described in Chapter 4) each recording file has been investigated manually and segmented into a set of smaller non-silent parts. The annotated data provide the information about the sound class existing in each labelled segment with its corresponding time-stamps. The 12 pre-defined annotation classes, such as bird vocalisations, wind noises, human speech and other background noises (see Table 4.2), were provided for the purpose of labelling the signal. In order to compare the detected outcome with the ground truth labels, apart from the 'bird vocalisations', all other labelling classes (including silences) in the annotated data are considered as the non-bird vocalisation class. Therefore, at the beginning, each manual annotated label file is transformed to a binary transcription signal (see Figure 5.4); where the one's value indicates bird vocalisation and the zero's value indicates non-bird vocalisation on each frame of the signal. This frame-based signal is named the "reference transcription signal" in this section.

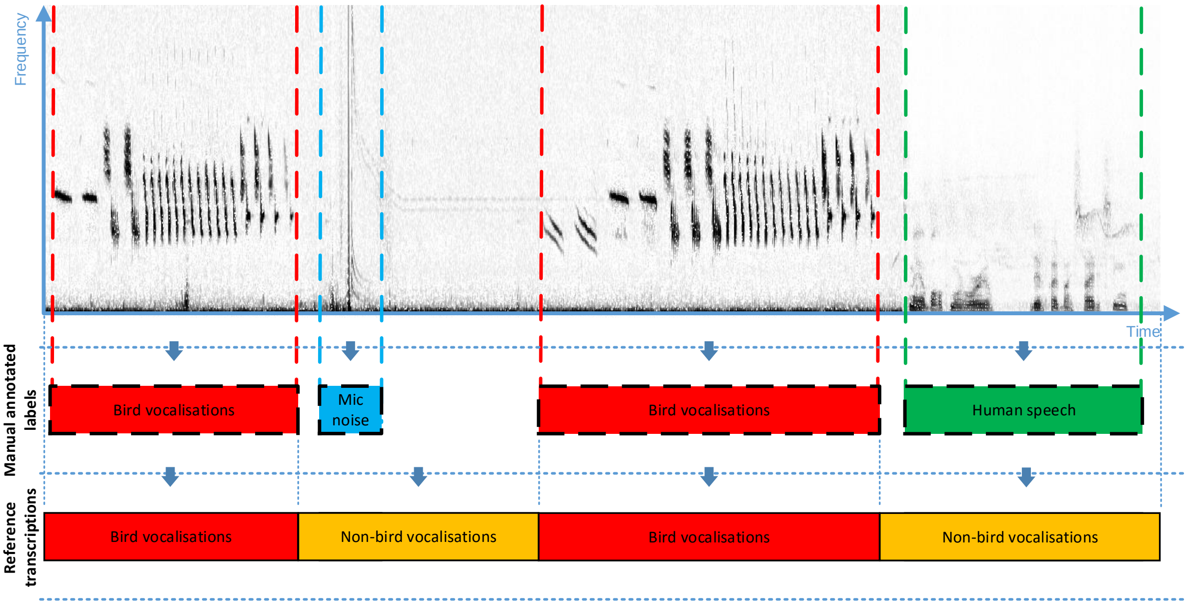


Figure 5.4: Obtaining the reference transcription signal from manual annotated labels

Fine tuning borders and splitting the (human labelling) bird singing segments into individual vocalisation

It is assumed that manual segmentation to mark the borders was not an entirely accurate process because of human listening errors. In addition, hand segmentation was not perfect enough to split all the bird vocalisation songs into smaller elements, similar to detected frequency tracks' segments. Therefore, an energy-based bird vocalisation detector is used with the purpose of fine tuning borders and possibly splitting the long bird singing segments into individual vocalisations.

The basic principles of this system are that first, it calculates the spectral energy from the input signal and then, since the energy of the bird singing parts are higher than the background energy, it compares these values with the threshold. The parts whose measured values are above the threshold will be considered as bird vocalisation and the rest will be considered as non-vocalisation. This energy-based detector system is described more in detail below and in Figure 5.5.

As Figure 5.5 illustrates, this procedure starts with taking the DFT at the frame level of the bird sound waveform along with calculating the energy at the spectral level (with the same parameter set-up presented in section 5.3.2). After obtaining the spectral energy for the entire sound signal, the energy sequence of each hand labelled segment (bird vocalisation segments only) is then passed to the energy-based bird vocalisation detector.

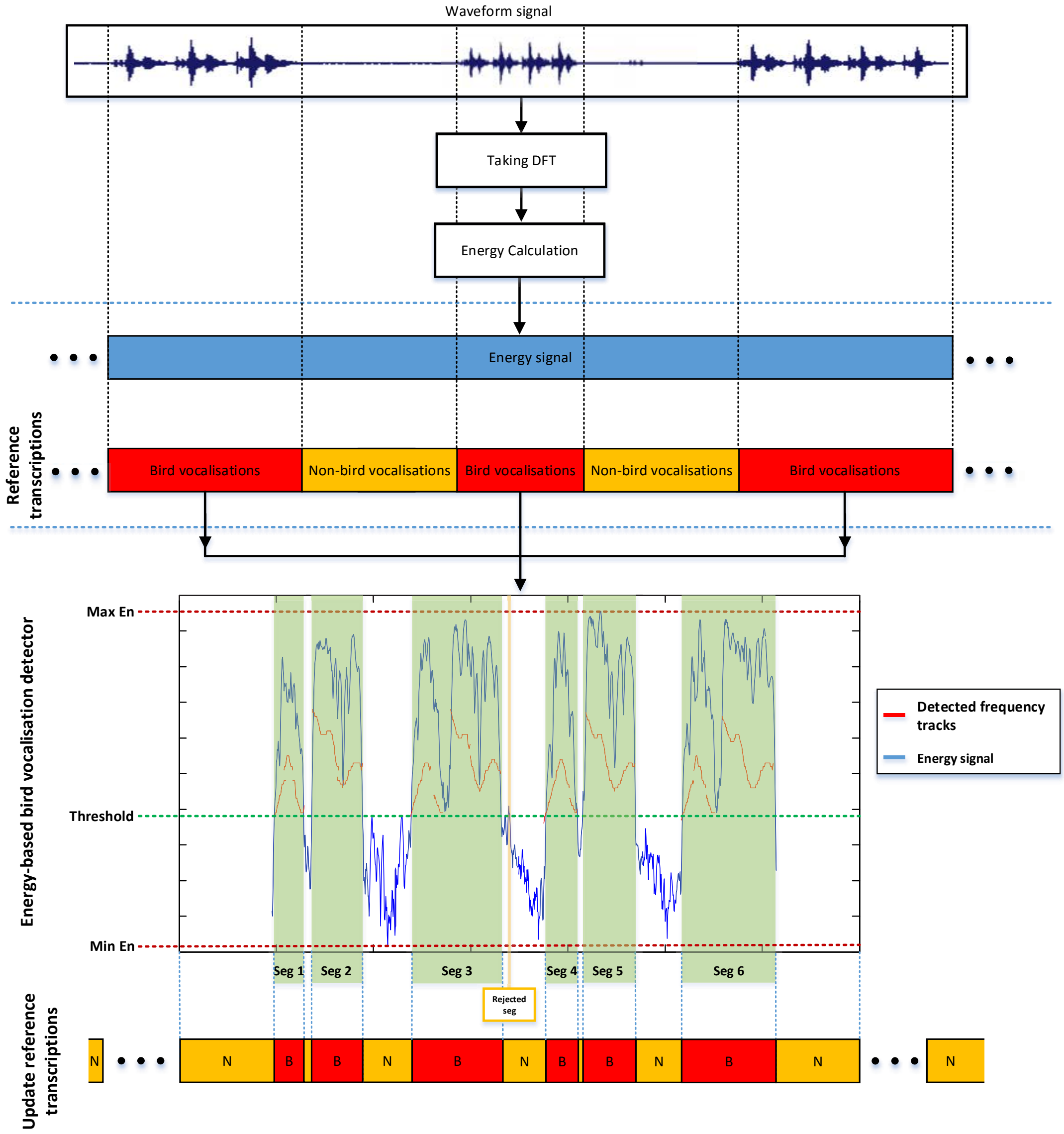


Figure 5.5: The procedure of fine tuning borders and splitting the bird vocalisation song into six individual vocalisation segments

It should be mentioned that the accuracy of the energy-based detector system depends heavily on the decision thresholds. Adaptation of the thresholds' value helps to track time-varying changes in the acoustic environments and hence, gives more reliable bird vocalisation segmentation results. The (adaptive) threshold for each incoming bird vocalisation segment is based on the energy levels of E_{min} and E_{max}, and it is estimated as:

$$Threshold = E_{min} + \alpha(E_{max} - E_{min}) \tag{5.1}$$

where E_{min} and E_{max} are the minimum and maximum energy values of the input energy frames, respectively; and α is a coefficient set to 0.4 empirically, as to achieve a very low false-acceptance error overall.

Along with the above threshold calculation, a stopping criterion is designed to check whether further classification is needed, as follows: if $E_{min} > (E_{max} - K)$, it is assumed that no gap or silent part was found within the entire input vocalisation segment; where K is a constant value and it is set to 30 dB empirically by observing the recordings.

Finally, the classification decision is based on the comparison of energy values of frames against the threshold's value. Bird vocalisation is declared if the measured values exceed the threshold; otherwise, that corresponding frame is classified as non-bird vocalisation. Meanwhile, the starting and ending points of the possible individual vocalisation segments were marked as the energy values crossed the threshold line.

Refinement of the reference transcription signals

As mentioned in section 5.3.3, the outcome of frequency track detection has been refined in terms of length and the average energy. In order to have a comparable condition, the same refinement rules are applied here as follows:

- Discard all the short bird vocalisation segments with a length less than 14 frames (e.g. rejected segment in Figure 5.5)

- Discard all the bird vocalisation segments where their average energy is 15 dB below the highest average segment energy in each recording.

The output of the energy-based detector is a binary decision on a frame-by-frame basis that presents the whole signal with a set of bird vocalisation and non-bird vocalisation frame-based labels. Finally, the reference transcription signal is amended based on the output of the energy-based detector.

5.4.4 Feature tracking evaluation measures

This section provides a performance analysis of the feature tracking procedure by comparing the outcome of the frequency tracks' extraction (detected output signal) with the reference transcription signal, in terms of bird vocalisation identification. These comparisons are assessed on a frame-by-frame basis and each frame of the detected output signal is checked with its corresponding frame (label) of the reference transcription signal. The performance measures of precision (the fraction of the identified instances that are relevant); recall (the fraction of relevant instances that are identified); and the false positive rate (FPR) are used to evaluate the performance of the above matchings as:

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{5.2}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{5.3}$$

$$FPR = \frac{FP}{FP + TN} \times 100\% \tag{5.4}$$

where, TP (True Positive) and TN (True Negative) are respectively, the number of frames correctly identified as belonging to the bird vocalisation class and the number of frames correctly identified as belonging to the non-bird class. In addition, FP (False Positive) and FN (False Negative) are respectively, the number of frames incorrectly identified as belonging to the non-bird vocalisation class and the number of frames that are not identified as belonging to the bird vocalisation class but should have been. The results for the performance analysis of the feature tracking procedure are shown in Table 5.2.

Table 5.2: Results for performance analysis of the bird vocalisation classification (precision and recall scores)

Name of bird species	Precision (%)	Recall (%)	FPR (%)
Bird 1: Carolina Wren	84.30	76.01	3.92
Bird 9: Louisiana Waterthrush	87.42	75.08	1.47
Bird 20: Hooded Warbler	73.80	77.88	1.87
Bird 31: Prothonotary Warbler	94.18	62.45	0.47
Bird 35: Kentucky Warbler	90.25	83.39	0.99
Bird 43: Savannah Sparrow	86.41	62.16	0.92
Overall:	86.06	72.83	1.61

5.4.5 Quantitative analysis of automatic segmentation procedure

The performance of the automatic segmentation procedure is assessed based on two criteria as follows:

1-Accuracy of segmentation procedure:

In this criterion, boundaries from the annotation data (reference transcriptions) are compared to the boundaries produced by the automatic algorithm. As can be seen in Figure 5.6, a simple rule is applied here for the entire comparison: if the boundary of the segment, which is produced by the automatic algorithm, falls within one reference boundary or has more than 50% overlap, that segment is considered as a hit or found boundary. In addition, insertions are produced when there are boundaries created by the automatic algorithm that do not match any annotated (reference) boundary. Deletions are produced when there is a boundary marked in the reference transcriptions, but the automatic detection algorithm produces no corresponding boundary. Once the alignment of two signals (reference vs detected output) was found, the quality of the automatic segmentation procedure can be obtained by percentage accuracy (PA) as:

$$PercentageAccuracy = \frac{N - D - I}{N} \times 100\% \tag{5.5}$$

where N is the total number of segments (boundaries) in the reference transcriptions; D and I are the total number of deletion and insertion errors, respectively.

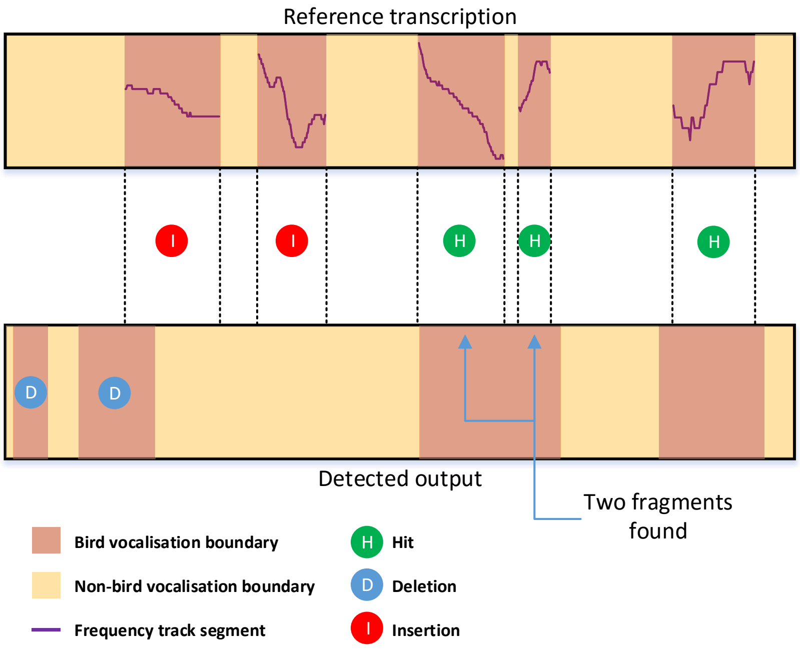


Figure 5.6: Examples of produced hits, insertions and deletions

2-Fragmentation score:

If there are several boundaries produced in the vicinity of only one reference boundary, fragmentation of the segment is considered for that corresponding boundary. In other words, if there is more than one hit available in one reference boundary, it is considered that the corresponding segment is split into more than one part in the automatic detection algorithm (see the example in Figure 5.6). The quality of the fragmentation can be presented in terms of the percentage score of the number of splits versus the total number of detected segments.

The results of the above segmentation analyses are presented in Table 5.3.

Table 5.3: Results for performance analysis of the segmentation task

Name of bird species	Percentage Accuracy (%)	Fragmentation score (%): split into			
		1 part	2 parts	3 parts	more than 3 parts
Bird 1: Carolina Wren	77.70	84.13	11.60	3.40	0.87
Bird 9: Louisiana Waterthrush	67.14	80.25	15.65	3.41	0.69
Bird 20: Hooded Warbler	61.64	66.49	19.63	8.78	5.09
Bird 31: Prothonotary Warbler	76.83	30.31	31.34	30.79	7.56
Bird 35: Kentucky Warbler	80.14	60.79	22.50	10.88	5.84
Bird 43: Savannah Sparrow	70.87	89.65	5.37	1.77	3.21
Overall:	72.39	68.60	17.68	9.84	3.88

5.4.6 Discussion on performance results

As can be seen in Table 5.2, the entire automatic detection system, when used at frame-level, can correctly detect over 72% of bird signal frames in the data, along with false-acceptance (FPR) of only 1.61% and very good overall precision of 86%. Furthermore, the obtained results for the performance analysis of the segmentation task (as can be seen in Table 5.3), show a very high level of overall accuracy of 72.39% for detecting the individual segments. Meanwhile, fragmentation scores demonstrate that some bird species have a complex structure in their vocalisations; e.g. about 70% of the found segments from the prothonotary warbler are split into more than two individual vocalisation parts (elements).

Here it has been observed that the improvement of the above evaluation results is heavily dependent on the manual data annotation procedure, since falsely labelled regions may be seriously detrimental to the detection accuracy. The manual inspection of examples of data showed that there were often several birds singing simultaneously, or other birds singing in background. The manual annotation requires a skilled level of expertise to precisely label these parts in their true label classes: as main bird vocalisation, or bird vocalisation in the background, with corresponding boundaries.

In general, it can be concluded that the proposed frequency tracks' detection method with the presented refinements steps, can be used to provide an accurate automatic segmentation of a bird signal recorded in a real natural environment, into individual vocalisation elements.

5.5 Summary

In this chapter, the segmentation and estimation of frequency tracks is performed, based on the dividing of the whole vocalisations' signal into distinct sinusoidal components, by employing the presented algorithm in [5, 6]. The detection procedure proceeds in two steps, the training and the recognition stages. First, the frame-based analysis has been performed in the training stage to detect the spectral peak points. Then, in the recognition stage, these obtained spectral points are classified into two sinusoidal and noise components, by comparing the corresponding magnitude shape and phase continuity distribution with the GMMs models. The experimental evaluation on the described vocalisation corpus successfully obtained a set of frequency tracks sequences for each recording file. Furthermore, performance analyses on the obtained results demonstrated that the entire frequency tracks' detection method provides very good accuracy in detecting the bird tonal spectral components, along with an accurate automatic segmentation of bird signals in real natural noisy environments.

Chapter 6

UNSUPERVISED DISCOVERY OF ACOUSTIC ELEMENTS IN BIRD VOCALISATIONS

6.1 Introduction

The segmentation and frequency track feature extraction procedure, as described in Chapter 5, provides a set of detected segments for a given audio recording, where they are only labelled with the name of the corresponding bird species. In general, there is not a wide range of publicly available annotated data for bird vocalisations. Such annotated bird acoustic data and the inventory of units of bird vocalisations are important for bioacousticians, for instance, to study differences between individuals and populations, or behaviour contexts; and furthermore for the development of more advanced automated systems for the processing of bird vocalisations [133].

The novel contribution of Chapter 6 is the development of a discovery approach in an unsupervised way that can find a set of individual vocalisation elements for each bird species, based on these detected frequency track segments. The resulting outcome of this novel approach also offers the label information for each discovered segment of the data.

This approach is accomplished in two consecutive stages. In the first stage (section 6.2), unlike a conventional dynamic time warping (DTW) algorithm which calculates the similarity of whole sequences, the presented modified DTW algorithm allows searching for multiple matches, possibly partial, between each pair of detected segments (frequency tracks segments). The outcome of the DTW search is a set of found partially matching paths, with their corresponding similarity values. These similarity values are called partial similarity scores, as they were obtained during the searching for partial paths. Each partial similarity score is calculated based on a combination of the cumulative distance of the DTW path match, the length of the matching path and the ratio of the length of the matching path to the total length of the segment. In the next stage (section 6.3.2), the outcome of the DTW searches is then used in an agglomerative hierarchical clustering approach to group all the homogeneous structured frequency tracks' segments into a set of distinct element-based vocalisation clusters. The final outcome is then used in Chapter 7, as the element-level vocalisation labelling information, to train the models in this proposed novel approach to the bird recognition system.

As mentioned earlier in Chapter 3 (section 3.4.2), elements can be taken as the smallest, structurally distinct, stereotyped acoustic units produced by a bird; and these can be thought of similarly as phonemes in the context of speech processing. Experimental results (section 6.4.3) show that the individual segments, which are grouped in each distinct vocalisation cluster, have similar structured patterns in terms of length and frequency values (see Figure 6.25). These found segments correspond to elements in a bird song signal and the patterns of the corresponding frequency values correspond to the pitch vocalisation pattern that was described in Chapter 3 (section 3.5.1). Therefore, in Chapter 6 (and later in Chapter 7), the term of “vocalisation pattern” refers to the pattern of the frequency values of the segments which are discovered in each distinct vocalisation cluster.

The following sections of this chapter explain the entire unsupervised discovery approach in more detail, along with the experimental evaluations’ section.

6.2 Segment similarity calculation using modified DTW

The application of the sinusoidal detection method described in Chapter 5, results in a form of an initial segmentation of the signal; for instance, signal frames in which no sinusoid is detected indicate no presence of tonal vocalisations. However, these initial frequency tracks segments may contain several repetitions of vocalisation elements and/or other tonal vocalisation sounds, which could be anywhere within the detected segments. Hence, to identify and group all the similar acoustic units together, a partial similarity searching method is required here.

In order to search for the partial and multiple matchings between a pair of detected segments, the use of conventional DTW that searches for the similarity of whole sequences is not suitable. Therefore, in this work a partial similarity searching method that employs modified DTW for a pair of segments to obtain the corresponding partial matching paths is introduced.

There are two main benefits of partial searching in DTW. First, this is effective for suppressing the result of potential wrong detection at the starting and ending point of segments. Secondly, it is useful for dealing with situations when the detected segment actually consists of several repetitions [134].

In the following sections, first the principal of a modified DTW method with employed constraints on the warping path is expressed; then, the whole proposed partial similarity searching method is presented. It is useful to note that the further stages of the processing are not dependent on the specific feature, so the presented work could also be applied to other applications with different types of features [133].

6.2.1 Modified DTW

For the purpose of simplicity, it can be considered that there are two single dimensional time series sequences as $X = (x_1, \cdots, x_{N_x})$ of length $N_x \in \mathbb{N}$ and $Y = (y_1, \cdots, y_{N_y})$ of length $N_y \in \mathbb{N}$. The distance matrix $d(x_i, y_j)$ is simply measured by using the Euclidean distance. The optimal warping path $w = (w_1, \cdots, w_K)$, defines a mapping between two sequences X and Y, in such a way that it minimises the cumulative distance $D_W(X, Y) = \sum_{k=1}^{K} d\left(x_{i_k}, y_{j_k}\right)$ [133].

Constraints on warping path:

In order to control the possible routes of the alignment path between two sequences and also to avoid any unpleasant the warping of the time-axis, a variety of rules and constraints are employed on warping path $w = (w_1, \cdots, w_K)$. These constraints are described as follows:

- Step pattern condition: This constraint allows the warping path to move one frame in horizontal, vertical or diagonal direction. As it can be seen in figure 1, the warping path w(k-1) is only one of the following $(i_k, j_k - 1)$, $(i_k - 1, j_k - 1)$ or $(i_k - 1, j_k)$. Additionally, this constraint satisfies the monotonicity and continuity conditions on the warping path as well.

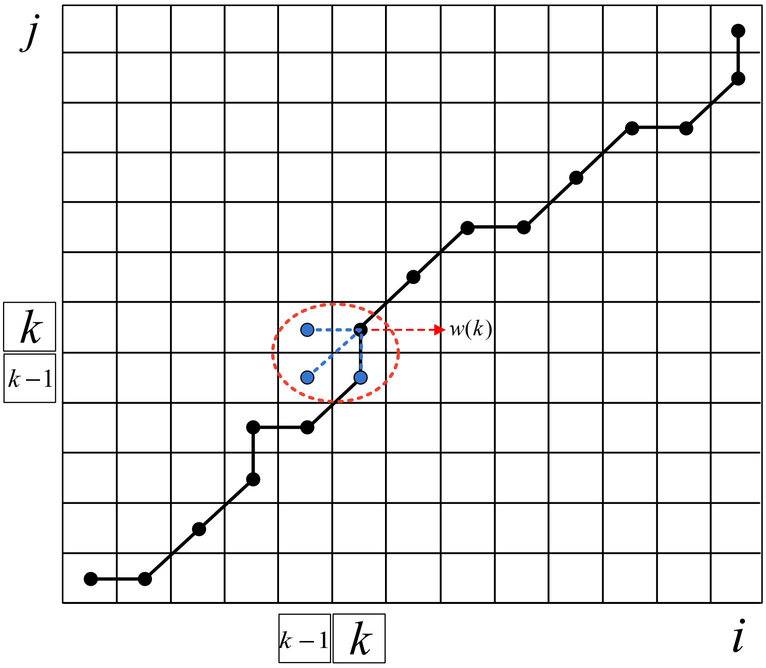


Figure 6.1: Illustration of warping path satisfying the step pattern condition

- Local constraints on time warping: This local constraint controls the possible relationship among several consecutive horizontal or vertical moves on the warping path. By setting up the allowance step parameter C, as it is set to 2 in this research, the modified DTW algorithm does not allow the path to have more than two consecutive moves in a horizontal or vertical direction (see Figure 6.2).

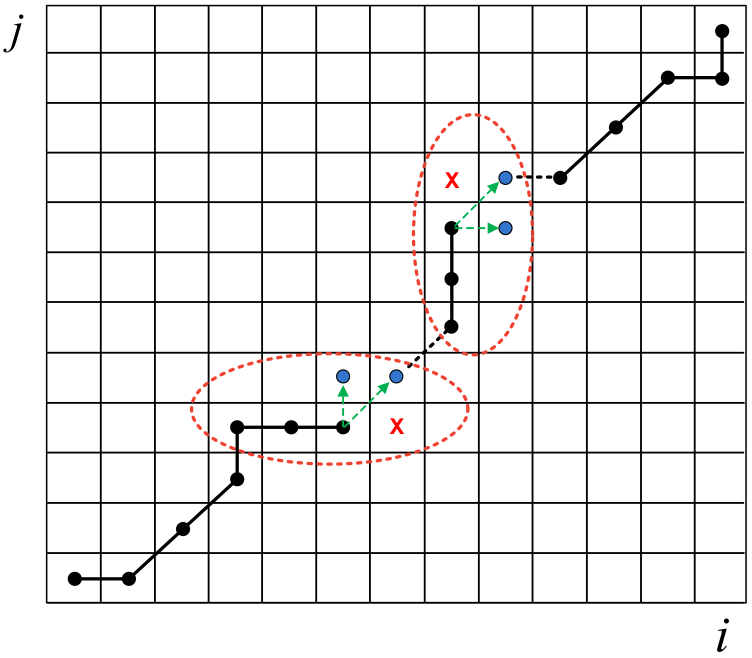


Figure 6.2: Illustration of warping path satisfying the local constraint conditions

- Boundary constraint: With conventional DTW the warping path usually starts at the bottom-left and ends at the top-tight of the matrix; whereas in modified a DTW different boundary constraint is employed, where the path starts from the first frame-time on segment Y and anywhere on segment X (see Figure 6.3). Additionally, this constraint allows the warping path w to end anywhere on the matrix. In other words, as it can be seen in Figure 6.3, each obtained cumulative path at the last frame-time on sequence Y, has a different starting point on sequence X and frame-time y_1.

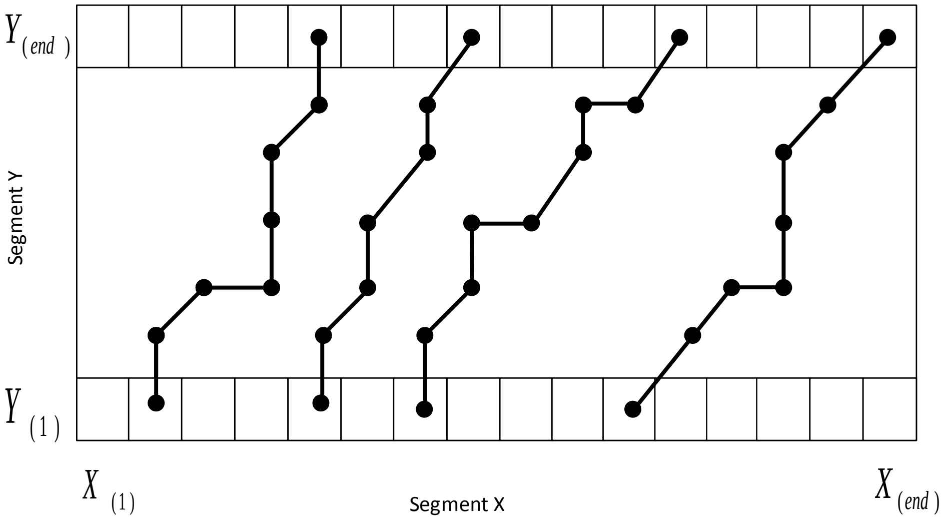


Figure 6.3: Illustration of warping path satisfying the boundary constraint

Normalising the cumulative distance value

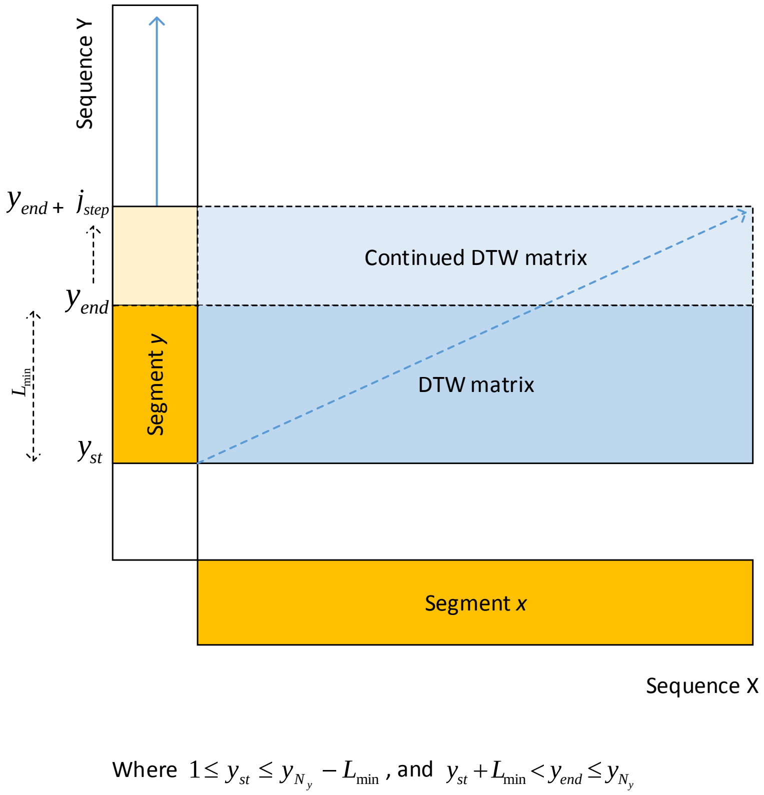
Since the distance $D_W(X, Y)$ accumulates over the warping path, the use of this distance value directly would cause the length of the sequence to affect the value. Thus, in modified DTW each cumulative distance value is normalised by the length of the warping path.

6.2.2 Partial and multiple matching using a modified DTW

In this section a partial searching method that employs the modified DTW algorithm to detect the similarity matches over two corresponding segments is presented.

Assuming again that there are two individual sequences: $X = (x_1, \cdots, x_{N_x})$, and $Y = (y_1, \cdots, y_{N_y})$; in order to find multiple partial matching paths, it is considered that the starting and ending points can be anywhere within the $N_X \times N_Y$ matrix. This can be done by performing several DTW searches in parallel, each considering a different starting point on one of the sequences, for example Y, and allowing the start anywhere on the other sequence X.

In other words, as it can be seen in Figure 6.4, during each searching procedure a range of frames of sequence Y is selected to compare with the entire sequence X. This shorter segment can be defined with two indices y_{st} and y_{end}, as starting and ending frame-time on sequence Y, respectively, where, $1 \leq y_{st} \leq y_{N_y}$, $1 \leq y_{end} \leq y_{N_y}$. In each DTW search iteration, the value of y_{st} does not change, whereas, the value of y_{end} would be increased by j_{step} number of frames up to frame-time y_{N_y} for further searching purposes. For clarity, this partial segment on the y-axis and the entire sequence of X are called segment y and x in this work, respectively.



Where $1 \leq y_{st} \leq y_{N_y} - L_{\min}$, and $y_{st} + L_{\min} < y_{end} \leq y_{N_y}$

Figure 6.4: Example of DTW search for a selected range of frames on sequence Y and the entire frames on sequence X.

Initial partial DTW searching procedure:

In the first part of DTW searching, $y_{st} = 1$ and $y_{end} = y_{st} + L_{min}$, where L_{min} is the minimum length of a cumulative path that the modified DTW algorithm considers for matching and this is used to discard short accidental match [133]. As mentioned previously, the boundary constraint in modified DTW, allows the cumulative path to start from the first frame-time on segment Y and anywhere on segment X; hence, the ending point of partial DTW paths can be anywhere on the x-axes. As the DTW calculation progresses, the cumulative distance values are obtained for subsequent points in the DTW matrix. Then, the values of the normalised cumulative distance can be examined at the frame-time y_{end} on the segment Y and any frame-time i on the sequence X. As the flowchart shows in Figure 6.5, there are two possible decisions at this stage as:

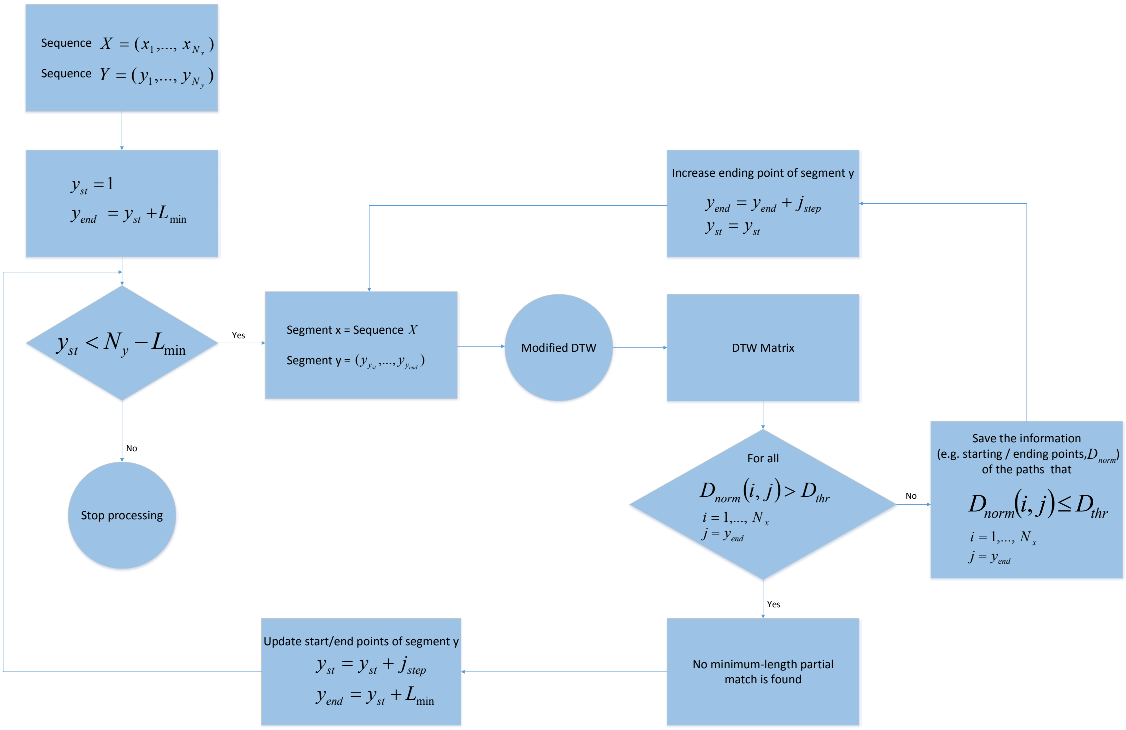

Sequence $X = (x_1,..., x_{N_x})$
Sequence $Y = (y_1,..., y_{N_y})$
$y_{st} = 1$
$y_{end} = y_{st} + L_{\min}$
Increase ending point of segment y
$y_{end} = y_{end} + j_{step}$
$y_{st} = y_{st}$
$y_{st} < N_y - L_{\min}$
Yes
Segment x = Sequence X
Segment y = $(y_{y_{st}},..., y_{y_{end}})$
Modified DTW
DTW Matrix
No
Stop processing
For all
$D_{norm}(i, j) > D_{thr}$
$i = 1,..., N_x$
$j = y_{end}$
No
Save the information
(e.g. starting / ending points,D_{norm})
of the paths that
$D_{norm}(i, j) \leq D_{thr}$
$i = 1,..., N_x$
$j = y_{end}$
Yes
Update start/end points of segment y
$y_{st} = y_{st} + j_{step}$
$y_{end} = y_{st} + L_{\min}$
No minimum-length partial match is found

Figure 6.5: DTW Partial searching procedure over two sequences X and Y

1- *Termination of current DTW calculation*: If there is no i such that the normalised cumulative distance D_{nrom} is below a given threshold D_{thr}, i.e. $D_{norm}(i, y_{end}) > D_{thr}$, for all $i = 1, \cdots, N_x$, then it is considered that there is no minimum-length partial match is found for the DTW search starting at the y_{st} on the sequence Y. Hence, the current DTW calculation of partial segment $y = (y_{y_{st}}, \cdots, y_{y_{end}})$ over the segment x (at the frame-time y_{st}) is stopped.

2- *The continuation of DTW searching on frame-time y_{st}*: If at least, there is such an i that $D_{norm}(i, y_{end}) \leq D_{thr}$, it is considered that the minimum partial match is found. There are three further steps available here. First, the algorithm saves the information about each of the found partial matching paths along with its associated starting and ending points, on both segments and the corresponding normalised cumulative distance. At the second step, in order to continue DTW searching at this starting frame-time y_{st}, the algorithm extends the length of partial segment y over sequence Y by updating its ending frame time as:

$$y_{end} = y_{end} + j_{step} \tag{6.1}$$

Where j_{step} is the minimum number of frames that is used to increase the length of segment y on sequence Y, after each successful DTW searching.

In the next step, the DTW calculation continues to obtain the cumulative distances for additional subsequent points in the DTW matrix. Once again, the values of normalised cumulative distances at the frame-time $y_{end} + j_{step}$ on the segment Y and any frame-time i on the segment X will be inspected, as means of the continuation DTW searching on the current frame-time y_{st}.

The three above steps: saving, inspecting and extending the ending point on segment y (changing the value of y_{end}), continue until no further match is found for any longer segment that starts at frame-time y_{st} on sequence Y.

For clarity purposes, Figure 6.6 shows an example of parallel DTW searching for two sequences X and Y. The segment y starts from frame-time y_{st} and ends at frame-time y_{end} on sequence Y. There are nine different warping paths among two time-frames y_{st} and y_{end} on segment y and anywhere on sequence X. The red circle on the DTW path shows that the value of the normalised cumulative distance is larger than a given threshold D_{thr} and the green circle shows vice versa. At the first level of DTW searching, at least six minimum-length partial matches are found as, B, D, E, F, G and I. During each level of searching, the algorithm always updates the associated information of all these successful matching paths. At the second level of searching, the DTW calculations continue to expand the cumulative distance matrix until frame time $y_{end} + j_{step}$ on the y-axis. By inspecting $D_{norm}(i, y_{end} + j_{step})$ for all $i = 1, \cdots, N_x$, it is considered that D_{norm} at the end of the path of E and F, are larger than D_{thr}. Hence, their associated ending points and its normalised distance remained as those that were obtained in the previous level of searching; whereas, the partial information in the other paths, B, D, G and I, were updated as the corresponding D_{norm} value is still below the threshold. The recursive DTW searching procedure continues for the current segment y and x, until the frame-time $y_{end} + 3j_{step}$ on the y-axis, where both values of the D_{norm} for paths D and G, are larger than the threshold. Hence there are no longer partial matchings path available for the current starting frame-time on sequence Y. Then, all the obtained partial matching paths, which are marked with yellow lines, with their associated partial information, will be presented in the DTW searching's outcome.

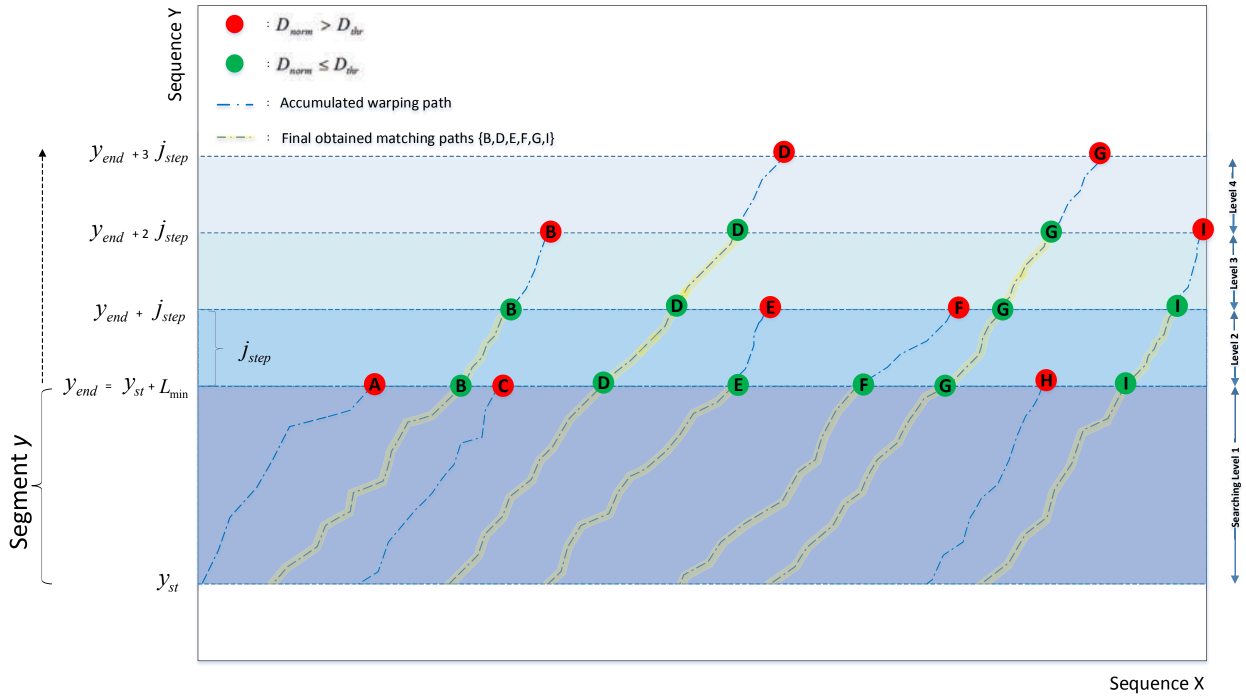


Figure 6.6: Illustration of the DTW searching procedure for candidate segments y over sequence X

Starting new partial DTW searching procedure on y-axis:

At the termination criterion of the above described steps, the algorithm starts a new DTW searching iteration for new partial segment y, by changing the starting and ending points on sequence Y as follows:

$$y_{st} = y_{st} + j_{step} \tag{6.2}$$

$$y_{end} = y_{st} + L_{min} \tag{6.3}$$

All the above described searching procedures are performed within new segment y and segment x, until the final stopping criterion is reached by the algorithm as:

$$y_{st} > N_y - L_{min} \tag{6.4}$$

where, N_y is the last frame-time on sequence Y.

6.2.3 Path selection refinement

So far, based on the above parallel DTW searchings along with two given sequences X and Y, a set of partial matching paths with their associated starting/ending points and the normalised cumulative distance are obtained. Typically, there are exponentially many matching paths available within the same rectangular area between the starting and ending points on two sequences. Thus, this stage presents a refinement method in such way that if there is more than one path falling within the rectangular area defined by the starting and ending points of the given partial warp path, only the path with maximum length is selected. This proposed refinement method is named a 'boxing' procedure in this thesis.

The boxing procedure employs a simple rule to discover the longest path as: if the start or end of a path, or any part of the path falls within a rectangular box of another path (vice versa), then keep only the longer one, otherwise keep both. This condition rule is then performed with several recursive procedures within the entire obtained partial paths of two corresponding segments.

For example, as Figure 6.7 shows, there are 10 obtained partial paths within sequences x and y, such as $P1$ to $P10$. The boxing procedure starts from the bottom-left and ends at the top of the matrix of x and y. First, the rectangular falling areas of two candidate paths $P1$ and $P2$ are checked. By finding the overlap area among them, $P2$ is selected as the longer path and $P1$ is discarded. In the next step (Figure 6.7-c), $P2$ and $P3$ are examined, followed by keeping $P3$ as the longer path. As the searching progresses until step i, other paths such as $P3$, $P5$, $P6$, $P7$ and $P8$ are discarded in competition with the longer path $P4$. As it can be seen in step j, there is no overlap between $P4$ and $P9$, thus the boxing procedure keeps both paths. Finally, the refinement procedure ends by removing $P10$ as a shorter path in competition with $P9$. Therefore, both paths of $P4$ and $P9$ remain as the results of the above boxing investigations.

6.2.4 Similarity calculation

For the purpose of further refinement on the obtained paths, the aim of this section is to select only one partial path, instead of many, within two subsequent segments. This can be done by keeping only the partial path which has the maximum similarity score within the others.

Up to this section, the similarity within each pair of segments is presented by the corresponding value of normalised cumulative distance. Instead of using that single distance criterion, a probability-like criterion that incorporates a combination of the normalised cumulative distance and the length of the matching path is used, to calculate the similarity score for each existing match. As a result, the similarity score is delivered from the below equation below as:

$$Sim(path_{i,j}) = P(x_D, x_L) \times l_R \tag{6.5}$$

where, $P(x_D, x_L)$ is the probability like criterion that can be presented by a specific logistic

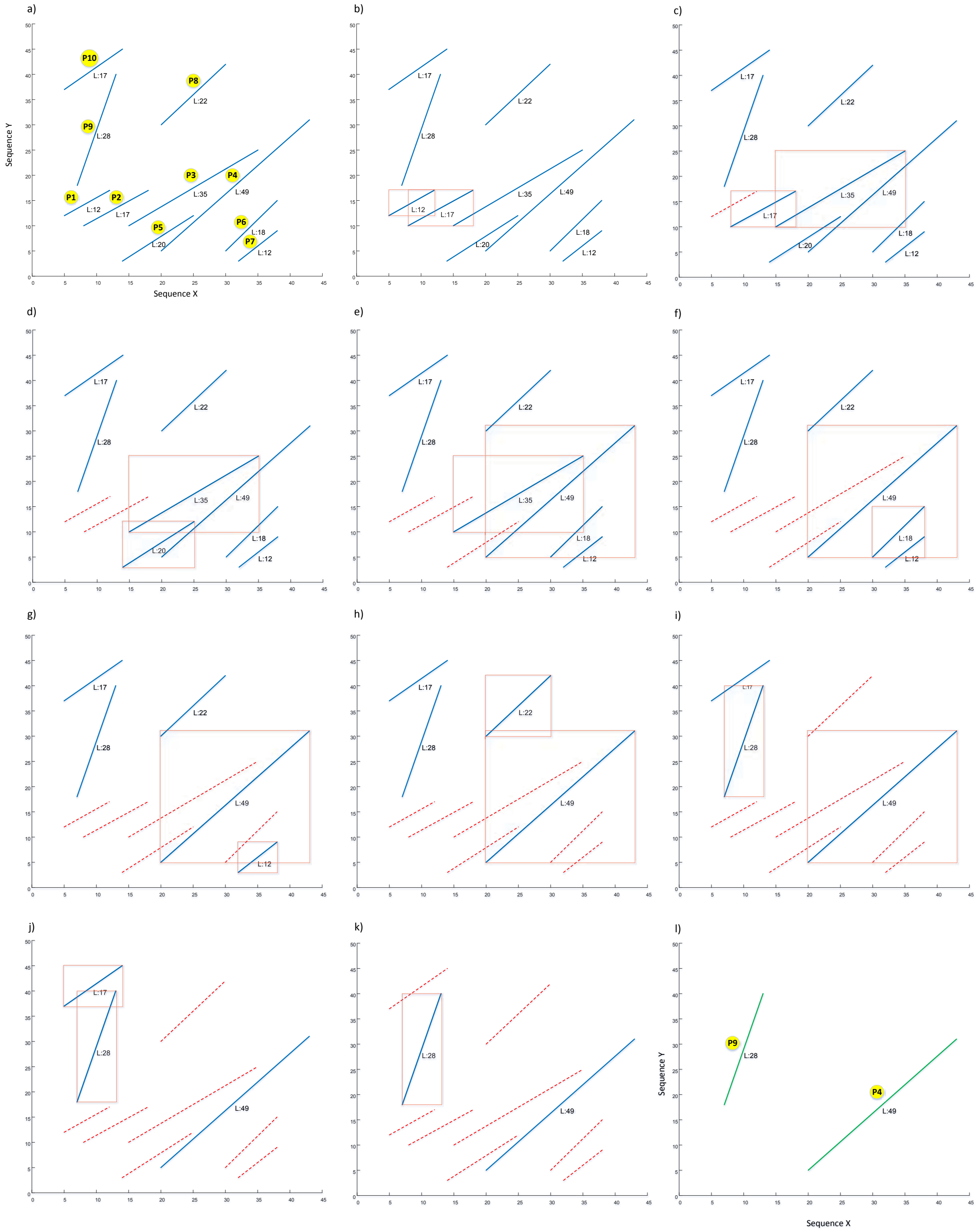


Figure 6.7: An Example of boxing procedure within ten partial matching paths, $P1, \cdots, P10$

function or sigmoid function, as:

$$P(x_D, x_L) = \frac{1}{1 + \exp(-\alpha_D(x_D - \beta_D))} \times \frac{1}{1 + \exp(-\alpha_L(x_L - \beta_L))} \tag{6.6}$$

The sigmoid function is a squashing function that produces an ‘S’ shape curve and the bound output $P(x_D, x_L)$ is always between 0 and 1; hence, it can be interpreted as a probability criterion function. The terms β_D and β_L are the midpoint’s constants and α_D and α_L are the steepness constants of the sigmoid curve, the x_D and x_L are the values of the normalised cumulative distance and the length of a matching path respectively. Figure 6.8 is an example of $P(x_D)$ that shows the effect of varying parameters’ midpoint and steepness constants.

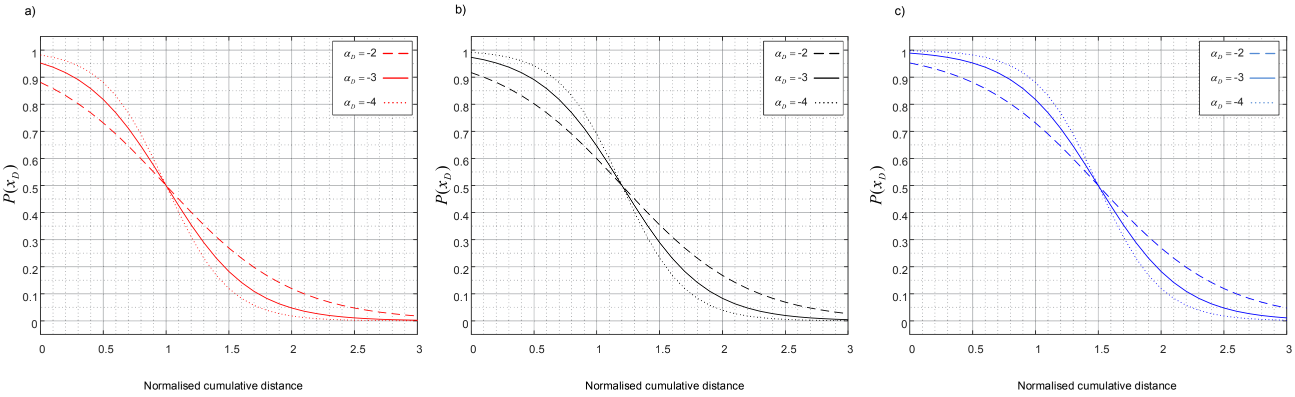


Figure 6.8: Effect of varying parameter: a) $\beta_D = 1$, b) $\beta_D = 1.2$, c) $\beta_D = 1.5$, in Sigmoid’s curve.

The term l_R in Equation 6.5, represents the ratio of the length of the matching $path_n$ to the total length of the corresponding segments i and j, as $L_{seg\ i}$ and $L_{seg\ j}$ respectively and it can be expressed as:

$$l_R = \sqrt{w_x(l_{R_x}) \times w_y(l_{R_y})} \tag{6.7}$$

$$l_{R_x} = \frac{l_x}{L_{seg\ i}} \times w_x,\ l_{R_y} = \frac{l_y}{L_{seg\ j}} \times w_y \tag{6.8}$$

where, l_x and l_y are the length of the scalar projection of $path_n$, on segment i and j

respectively (see figure 6.9); w_x and w_y are the weight coefficients and can be obtained as (see figure 6.10):

$$w = \begin{cases} 1.8(l_R) - 0.8 & , l_R \geq 0.5 \\ 0.1 & , l_R < 0.5 \end{cases} \tag{6.9}$$

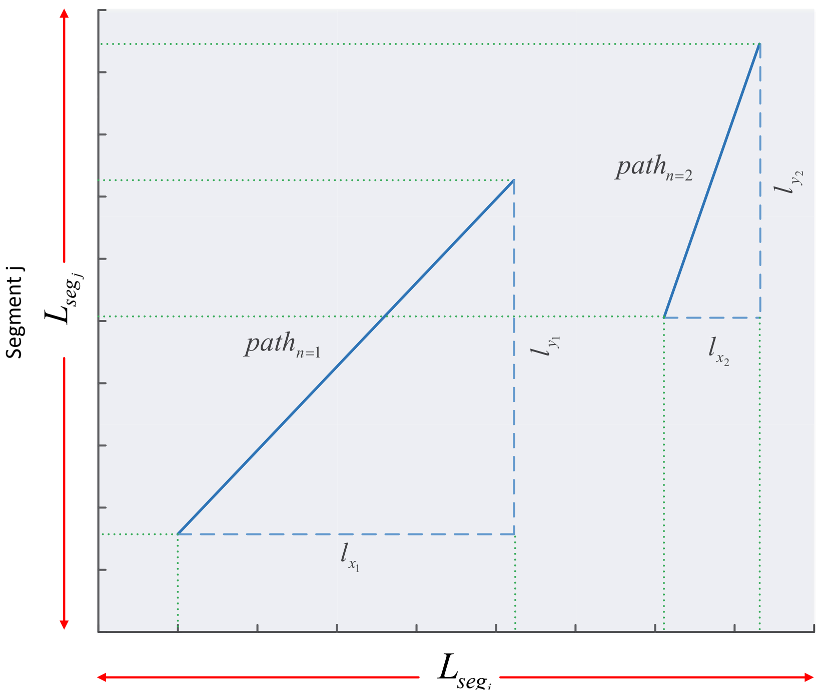


Figure 6.9: An example of the scalar projections of matching paths on given segments i and j

By keeping the path which has the highest similarity score within the others, the corresponding matching path between segments i and j is described as follows: start and end points of segment i, start and end points of segment j and the similarity score $sim_{(i,j)}$ (see Figure 6.11).

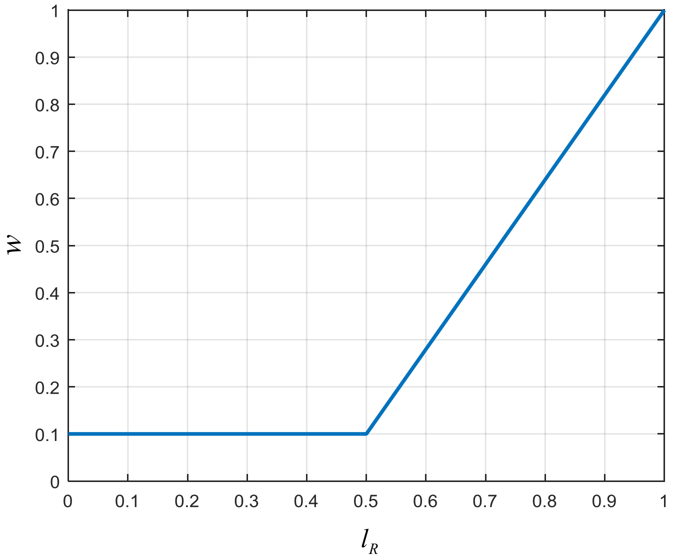


Figure 6.10: An example of the curve for weight function $w(l_R)$, for $0 < l_R < 1$

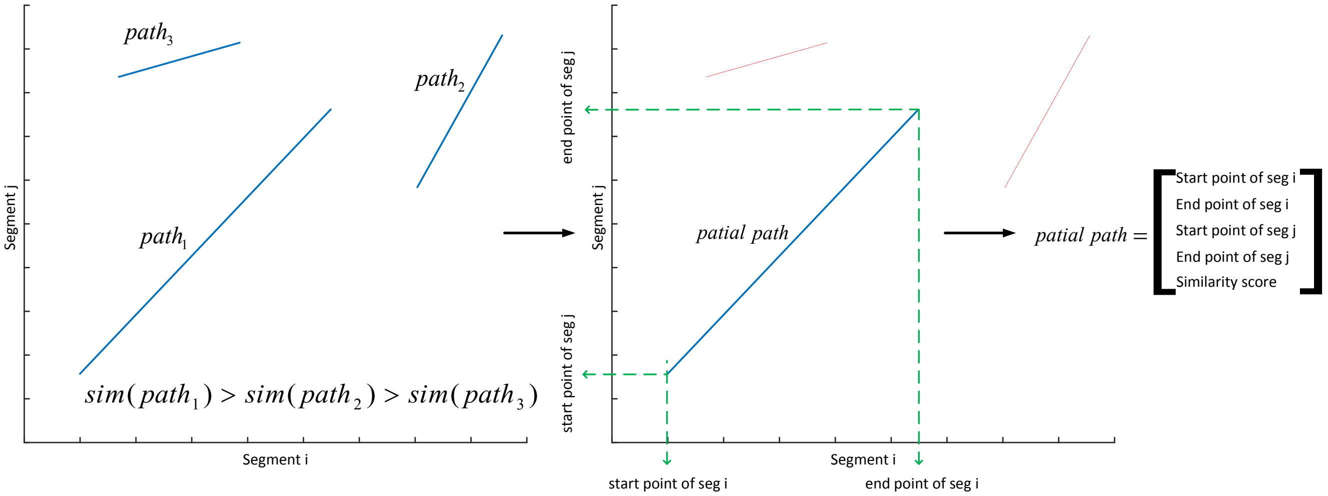


Figure 6.11: An illustration of found partial path between segments i and j

6.2.5 Structure of DTW output results

After the processing of all the above partial similarity searching procedure for N number of given sequences, the entire result, including the information of the obtained partial path and the similarity score for all the segment pairs, can be presented by two $(N \times N)$ information matrices: the partial path matrix and the similarity distance matrix.

1- Similarity distance matrix: a symmetric square matrix, where each cell of the matrix shows the similarity value for a pair of segments at i_{th} row and j_{th} column. For example, as Figure 6.12(a) illustrates the $simDisMat(m, n)$ shows the found similarity score between two segments m and n equal to 0.98.

2- Partial path matrix: this is also a symmetric square matrix where each cell of the matrix shows the starting/ending information of a detected path for a pair of segments at i_{th} row and j_{th} column. For example, as it can be seen in Figure 6.12(b), the obtained path within segments m and n is defined as;

$$partPathMat(m, n) = \begin{bmatrix} Start\ of segment\ m \\ \\ End\ of segment\ m \\ \\ Start\ of segment\ n \\ \\ End\ of segment\ n \end{bmatrix} = \begin{bmatrix} 1 \\ \\ 40 \\ \\ 3 \\ \\ 42 \end{bmatrix} \tag{6.10}$$

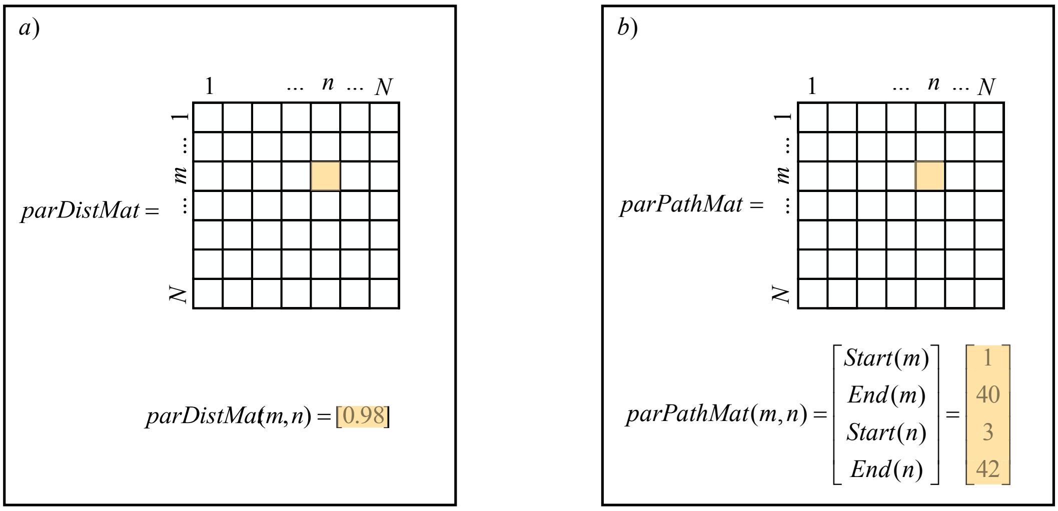


Figure 6.12: The result of partial match searching for given N numbers of sequences can be expressed by two matrices a) partial distance matrix and b) partial path matrix.

An example of the result obtained by the whole DTW searching procedure on a pair of real-world bird recordings is given in Figure 6.12. The blue lines indicate all the partial matchings found. For simplicity, the lines are drawn by connecting the starting and ending points of the found match. It can be seen that the procedure found 13 partial matches between the given two sequences that align well to each other.

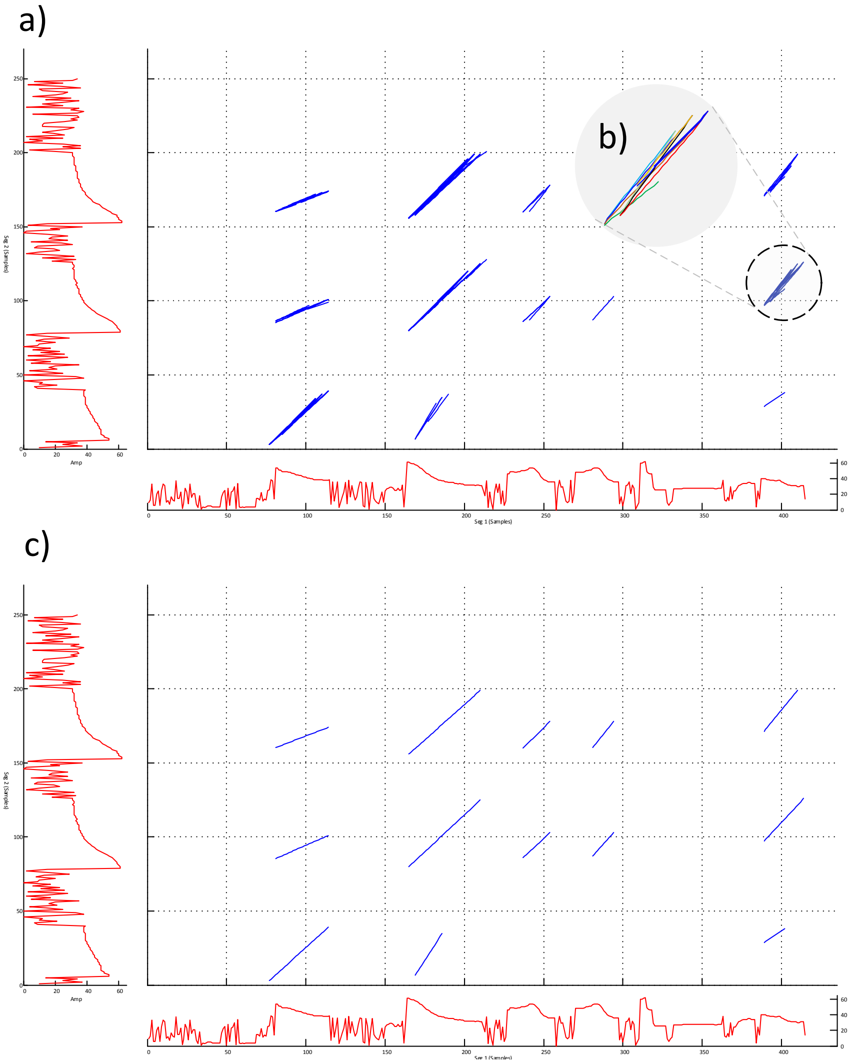


Figure 6.13: Illustration of the output of multiple partial matchings, a) before path selection refinement step, b) within an example boxing procedure, c) after similarity calculation step, for two given bird recordings.

6.3 Clustering

The outcome of the partial DTW algorithm is a large collection of segments and their associated distances one to another. Here, a modified agglomerative hierarchical clustering method is presented to identify and group together all structurally similar segments that were produced by a particular bird.

6.3.1 Problems and limitations with conventional hierarchical clustering algorithms

Although the conventional hierarchical clustering algorithm is thought for as simple, it often comes upon some difficulties regarding the selection of merge links. In the ordinary techniques, including the single linkage and the group average method, the individuals or small groups have a tendency to cluster together at relatively early levels [45]. Also, once the clusters at one level are merged into another group of individuals, the further merging at the next step is on the newly formed cluster. Hence, the merge decision is such a critical performance, as there is no undo or swapping procedure available in this method [136]. If the merge or split decisions are not taken well, the whole clustering procedure may lead to a poor performance result. Moreover, as the ordinary method does not clearly evaluate the merging or splitting procedure [136], a few internal pre-examinations are needed to analyse the further structure of the clusters just before taking any merging or splitting decisions.

One promising guideline for enhancing the clustering quality of the hierarchical methods is to incorporate the external relationship information data alongside of the distance matrix; such as partial matching information in this research that describes the relations within whole entities. Furthermore, at the end of each merging iteration, the mutual relationship information of each cluster, should be updated in parallel with distances among the clusters. Therefore, the traditional approaches are often unable to deal with this information [137].

6.3.2 Modified unsupervised clustering method (Cluster representative calculation)

In this section, a modified agglomerative clustering method that employs several specific rules and constraints is presented, in order to identify and group together all structurally similar segments that were produced by a particular bird. As such, this algorithm incorporates both obtained similarity distance and partial relationship information matrices, as the input attributes of clustering.

1-Initial step and parameters' definitions:

As mentioned in section 6.2.5, the output of DTW searching for N number of given segments of frequency tracks $seg_1, \cdots, seg_N$, is represented by two matrices: the similarity distance matrix and the partial path matrix, where both data describe the partial intrinsic characteristic along with each indices. In the proposed clustering method, $simDisMat_{(N,N)}$ refers to the similarity distance matrix and it is considered as the major clustering data source or distances; whereas the partial path matrix or $parPathMat_{(N,N)}$, is incorporated as the external relationship information matrix to influence the grouping decisions.

At the initial stage, each frequency tracks segment is assumed as a distinct single-member group or $G_i = seg_i$, where $i = 1, \cdots, N$. Then the clustering proceeds with several consecutive iterations, where the initial aim of each iteration is to find the two closest segments that have the largest pairwise similarity score within all the rows and columns of *simDisMat*. The resulting pair of segments at each level is presented by seg_m and seg_n in such a way that,

$$simDisMat(m, n) = max(simDisMat) \tag{6.11}$$

In the first level of clustering, there is no restriction or condition to stop the merging, therefore the first segment m and n are merged into a larger group of $G_{(m,n)} = \{seg_m, seg_n\}$.

Despite the ordinary hierarchical clustering methods where the pairwise distances of the new formed cluster and the remaining clusters are often measured after each merging step, this method does not calculate these pairwise similarities. Instead, the algorithm calculates the corresponding group average similarity or *groupAveSim* for each cluster with multiple individuals. The group average similarity for cluster G with n number of individuals can be estimated as:

$$groupAveSim_{G=seg_1,\cdots,seg_n} = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n} simDisMat(i, j)}{\frac{n^2-n}{2}} \tag{6.12}$$

For example, if $n = 3$ the group average similarity for cluster $G = \{seg_1, seg_2, seg_3\}$ can be evaluated as;

$$groupAveSim_G = \frac{simDisMat(1, 2) + simDisMat(1, 3) + simDisMat(2, 3)}{3} \tag{6.13}$$

The group average similarity of each cluster with two individuals, including the first formed cluster $G_{\{m,n\}} = \{seg_m, seg_n\}$, is equal to the pairwise similarity between segment m and n as *simDisMat(m,n).*

In addition, the mutual path overlap, or *overlap*, of both segments of any two-member clusters, e.g. segment m and n in cluster $G_{\{m,n\}} = \{seg_m, seg_n\}$, are obtained from *parPathMat*, directly after the joining of two segments as follows;

$$partPathMat(m,n) = \begin{bmatrix} Start\ of segment\ m \\ End\ of segment\ m \\ Start\ of segment\ n \\ End\ of segment\ n \end{bmatrix} = \begin{bmatrix} 1 \\ 40 \\ 3 \\ 42 \end{bmatrix} \tag{6.14}$$

$$overlap_m = \begin{bmatrix} Start\ of segment\ m \\ End\ of segment\ m \end{bmatrix} = \begin{bmatrix} 1 \\ 40 \end{bmatrix} \tag{6.15}$$

$$overlap_n = \begin{bmatrix} Start\ of segment\ n \\ End\ of segment\ n \end{bmatrix} = \begin{bmatrix} 3 \\ 42 \end{bmatrix} \tag{6.16}$$

As the *overlap* parameter shows the partial matching boundaries of each segment with entire linked segments in its associated group, this mutual overlap area of the segment can be changed or shortened only when another individual segment or a multi-member cluster is joined into the same cluster where the corresponding segments are available there. In both situations, the *overlap* parameter of all the segments of the new larger formed cluster, should be updated. The updating procedure of theses situations will be discussed later in this chapter.

Both the mentioned clustering parameters, *groupAveSim* and *overlap*, alongside with the partial matching path information matrix, can describe the likeness structure of each formed group and they will be used as a means of pre-joining the evaluations to analyse the next merge decisions, for the rest of the clustering procedure.

2-Rules of joining:

As the clustering method progresses to the second level, a set of rules and restrictions are employed to influence the possible group merging transaction. In other words, the merging into groups only happen when the joining criterion is reached by the algorithm.

At the beginning of each iteration, as mentioned previously, the algorithm searches for the next closest pair of segments. After that, both obtained segments m and n, which have the next largest pairwise similarity value in the distance matrix, are examined in terms of whether they are a member of any obtained group. The result of this examination leads the algorithm to one of the three following stages:

2.1- Both segments are chosen from the remaining single segments:

If both candidate segments are of the single-type cluster, or an individual segment, there is still no restriction available here to stop the merging procedure. Thus, these two segments are joined into the new cluster G with an occupancy of two; followed by the calculation of the corresponding group's average similarity value or *groupAveSim* and the mutual path's overlaps, or *overlap*, for both segments m and n (see Figure 6.14).

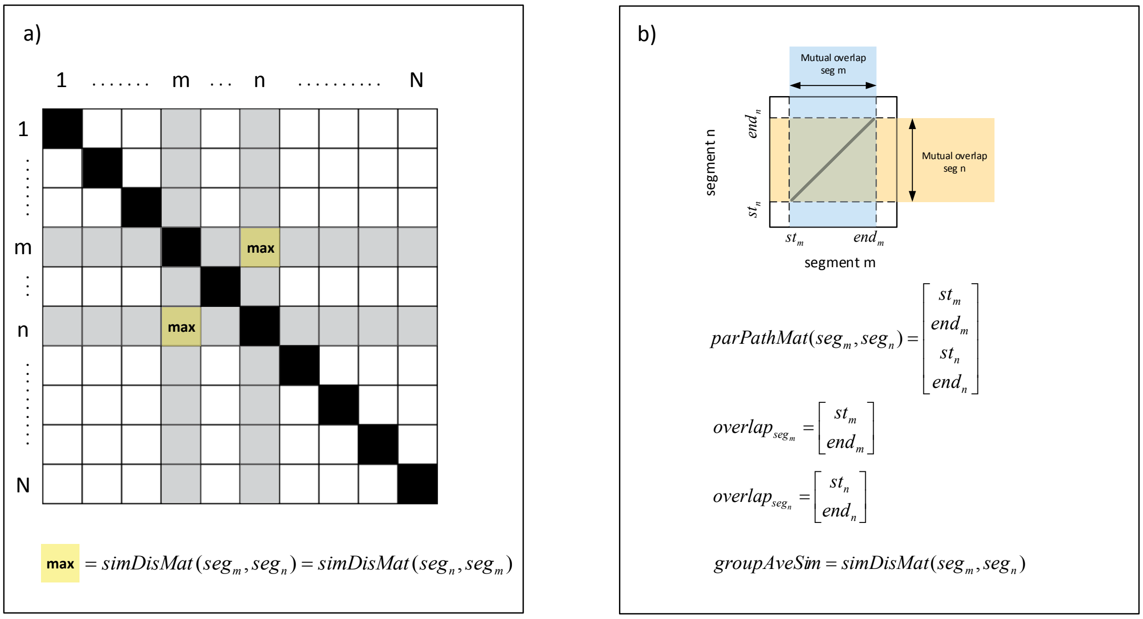

a)
1 .... m ... n .... N
1
m
n
N
max
max
max $= simDisMat(seg_m, seg_n) = simDisMat(seg_n, seg_m)$
b)
Mutual overlap seg m
Mutual overlap seg n
segment n
end_n
st_n
st_m
end_m
segment m
$parPathMat(seg_m, seg_n) = \begin{bmatrix} st_m \\ end_m \\ st_n \\ end_n \end{bmatrix}$
$overlap_{seg_m} = \begin{bmatrix} st_m \\ end_m \end{bmatrix}$
$overlap_{seg_n} = \begin{bmatrix} st_n \\ end_n \end{bmatrix}$
$groupAveSim = simDisMat(seg_m, seg_n)$

Figure 6.14: a) Searching the maximum pairwise similarity within distance matrix. b) Obtaining the group average similarity value and the mutual path overlaps of the new formed group of two individual segments m and n.

2.2- The first segment is a single-typed cluster and the second one is associated to a specific group with several individuals, or vice versa:

If one of the candidate segments is previously linked to any group with several individuals, the algorithm does not allow the merging to proceed until all the further conditions are checked. These conditions with corresponding evaluations at this stage are described as below:

2.2.1- Checking the mutual path overlaps with obtained partial matching path:

For clarity purposes, assume that the seg_m is a single segment and seg_n is one of the current segments of $Group\ A\ \{seg_n, seg_a, seg_b, seg_c\}$. At this stage, the current mutual path overlap of each individual member of group A, including segment n, is examined with the partial match of the corresponding member that the path between segments m and the corresponding segment shows. As it can be seen in Figure 6.15, the information of the partial path between single segment m and each member of group A can be delivered from its subsequent partial information matrix $parPathMat$ as:

$$parPathMat(seg_m, seg_i) = \begin{bmatrix} Start\ of segment\ m \\ End\ of segment\ m \\ Start\ of segment\ i \\ End\ of segment\ i \end{bmatrix} \tag{6.17}$$

Where, segment i is the corresponding member of group A as $i \in \{n, a, b, c\}$; the starting and ending points of segment i defines the obtained matching part of segment i.

If there is such an overlap found for all pairwise examinations, then the algorithm goes to the next checking point. If not, the algorithm does not allow the segment m to join into a larger cluster A, followed by continuing with a new clustering iteration to find the next highest similarity and the next closest admission pair of segments.

Figure 6.15 displays both successful and failed examples, alongside the entire overlap checking procedure between segment m and all segments in group A.

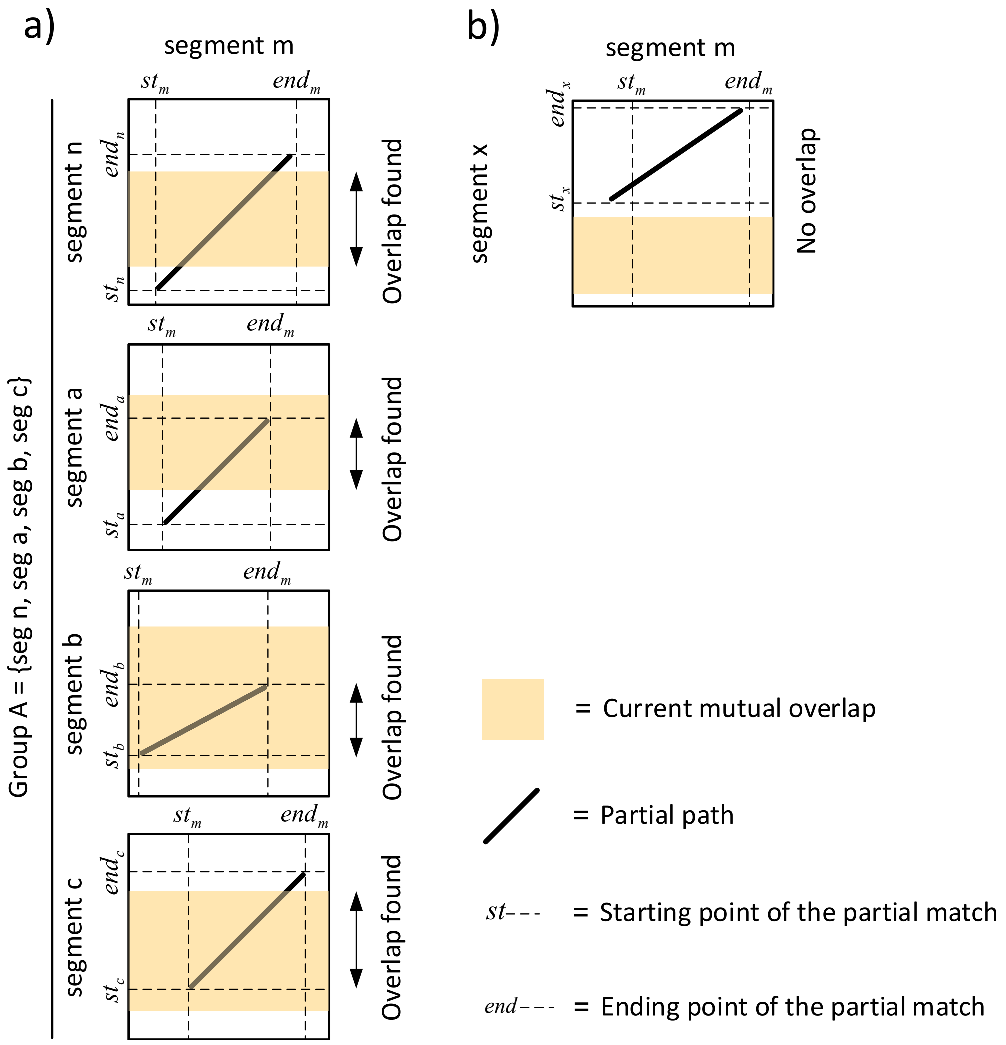


Figure 6.15: Overlap searching between partial match and mutual overlap paths: a) shows all the successful checking procedures between segment m and all segments in group A. b) shows an example of failed overlap searching.

2.2.2- Checking the drop in the group average similarity of the possible merge with candidate segment, with the given threshold thr_{drop}*:*

Clusters with low occupancy are formed at the early stages of clustering. As the method progresses, the additional individuals or other groups may change the mutual structure of these clusters rapidly at the initial stages. In order to prevent such fast disorders, the adopted criterion for judging when the candidate segment (or group; as this will be described later) should not be added to the current cluster is used.

The current check point analyses the prospective likeness structure of the group with the additional single segment, by calculating the drop in the group average similarity. As it is described previously, the group similarity is calculated averagely, since it has been formed from the merge of two single segments. Therefore, the drop for any group G can be obtained as:

$$Drop = groupAveSim_{\hat{G}} - goupAveSim_{G} \tag{6.18}$$

where, G is the group with current occupancy before merging as $G = \{seg_1, \cdots, seg_n\}$, $\hat{G}$ is the group with addition candidate segment seg_{n+1} as $\hat{G} = \{seg_1, \cdots, seg_{n+1}\}$.

If the drop is greater than a given threshold thr_{drop}, i.e., $Drop > thr_{drop}$, then the candidate segment should not be added to the cluster G. If it is not greater, the final joining criterion is reached, thus the algorithm allows the candidate segment, i.e., segment m, to merge into the corresponding group G, where the other obtained closest segment, i.e. segment n, is associated.

In the next step, as the joining procedure is accomplished, the new group average similarity and all the corresponding mutual overlap paths should update in the algorithm. Therefore, the value of the calculated group similarity of $\hat{G}$ or $groupAveSim_{\hat{G}}$, is placed into the group similarity value of the updated cluster G, with the additional candidate segment or $groupAveSim_{G}$.

Moreover, as it can be seen in Figure 6.16, the subsequent mutual overlap path of each segment in new cluster G, i.e. seg_i for all $i = \{n, a, b, c\}$, excluding the newly merged segment m, can be replaced by the corresponding found overlap as follows:

$$overlap_{seg_i} = overlap_{seg_i} \ \cap \ [st_{seg_i} : end_{seg_i}] \tag{6.19}$$

where st_{seg_i} and end_{seg_i} are delivered from:

$$parPathMat(seg_m, seg_i) = \begin{bmatrix} st_{seg_m} \\ end_{seg_m} \\ st_{seg_i} \\ end_{seg_i} \end{bmatrix} \tag{6.20}$$

Also, the mutual overlap path for the newly merged individual, i.e. segment m, can be obtained as:

$$overlap_{seg_m} = \begin{bmatrix} max(st_{seg_m}) \\ min(end_{seg_m}) \end{bmatrix} \tag{6.21}$$

where, $max(st_{seg_m})$ takes the largest value of st_{seg_m} in $parPathMat(seg_m, seg_i)$ for all $i = \{n, a, b, c\}$, and $min(end_{seg_m})$ takes the smallest value of end_{seg_m} in $parPathMat(seg_m, seg_i)$ for all $i = \{n, a, b, c\}$.

Finally, before the new clustering iteration begins, all the pairwise similarities within all the members of the newly formed cluster are assumed as zero in the distance matrix for the purpose of finding the next admissions

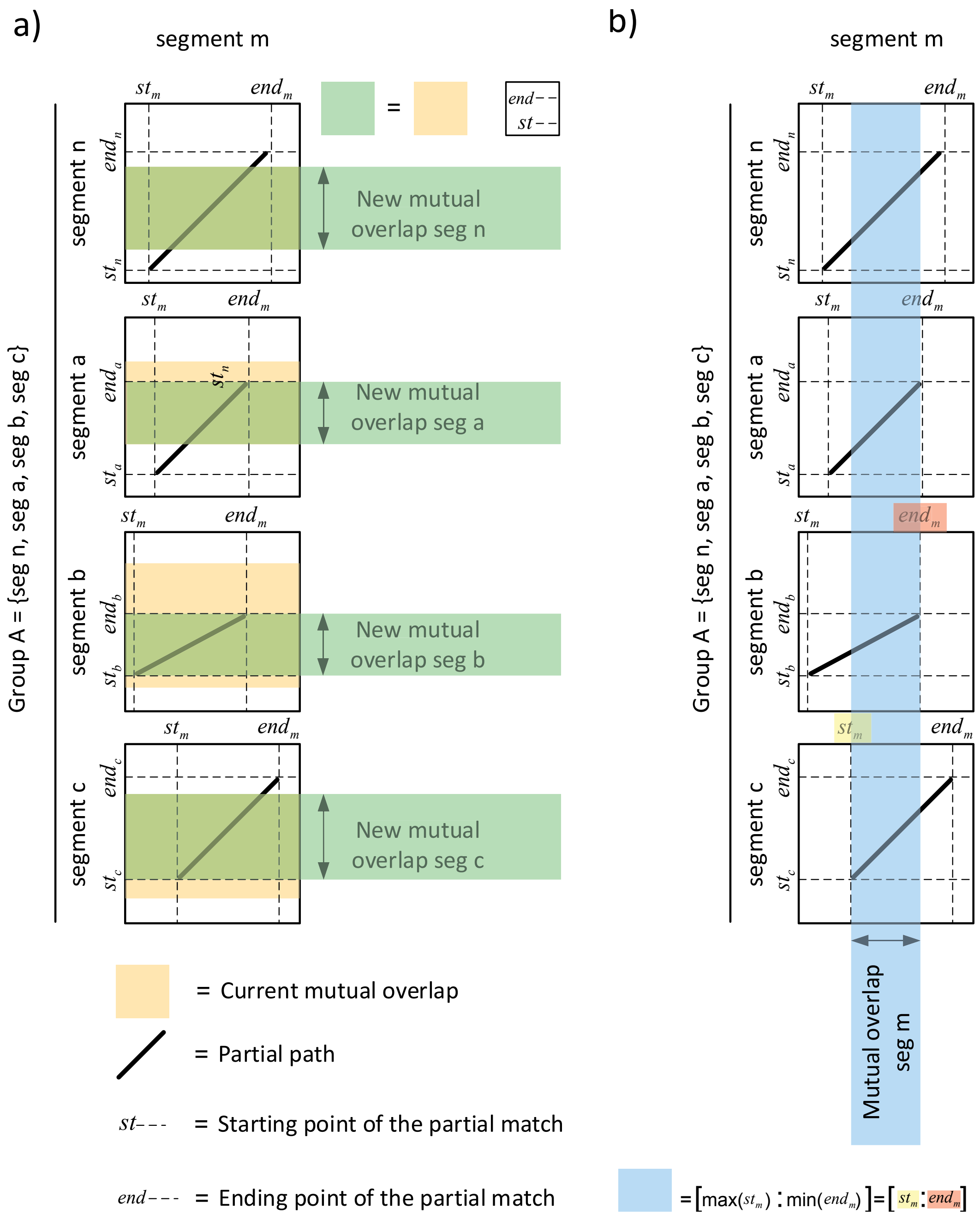


Figure 6.16: Updating the merge parameters: a) shows the updating procedure of the new mutual overlap path for all the old members of group A, such as segments n, a, b and c. b) shows the obtaining procedure of the mutual overlap path for newly merged segment m.

2.3- Both segments are previously assigned into the two different clusters with several individuals:

at this stage, as both the candidates belong to the two various multi-member clusters, the merge would be within these two clusters with all their associated segments, only if the corresponding merge criterion is reached. For simplicity purposes, assume that segment m and n are the current closest pair of the segment, where $seg_m \in Group_A$, $seg_n \in Group_B$, $Group_A = \{seg_m, seg_d, seg_e\}$ and $Group_B = \{seg_n, seg_a, seg_b, seg_c\}$.

Initially, the algorithm does not allow the merge to proceed, between $Group_A$ and $Group_B$, until all the further conditions are checked. The principles of these step conditions are similar to the above described conditions in the previous section, only with some additional evaluations within the entire candidate clusters.

2.3.1- Checking the mutual path overlap of the entire segments of both candidate groups:

In contrast with the previous section where the comparisons were only between the single segment, i.e. segment m, and other members in $Group_B$, at this check point the comparisons are between the entire segments of both groups A and B. In other words, at each examination, the current mutual path overlap of each individual member of $Group_B$ including seg_n, seg_a, seg_b and seg_c, is examined with its corresponding partial path, which is obtained with each segment of $Group_A$, including seg_m, seg_d and seg_a (see Figure 6.18), and vice versa (see Figure 6.17).

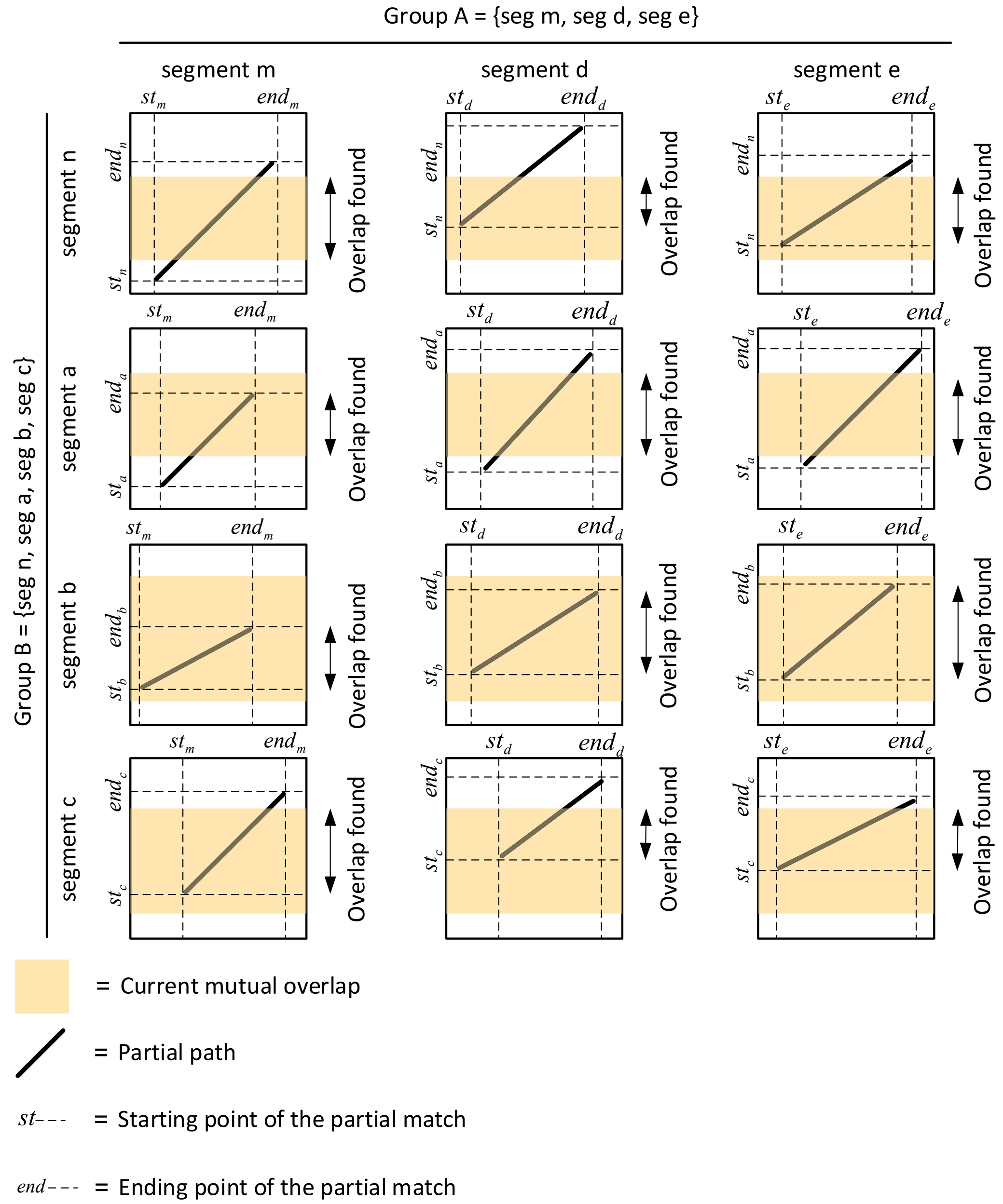


Figure 6.17: Searching for overlap between the mutual overlap path of each member of group B and its corresponding partial path with group A.

If there is such an overlap found for all the examinations, then the algorithm goes to the next checking point; if not, the algorithm does not allow the merging between these two candidate groups, it then proceeds to a new clustering iteration to find the next highest similarity and the next closest admission pair of segments.

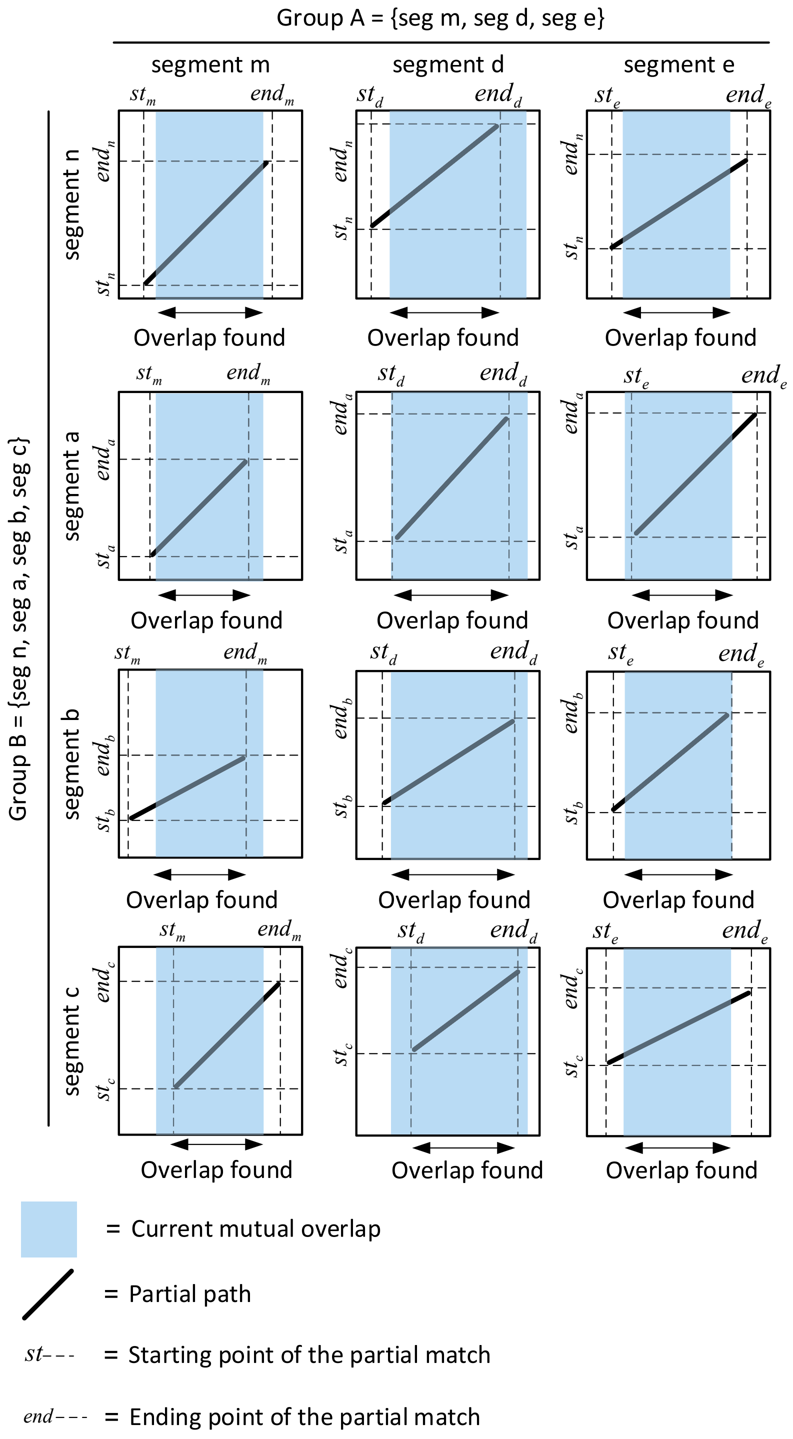


Figure 6.18: Searching for overlap between the mutual overlap path of each member of group A and its corresponding partial path with group B.

2.3.2- Checking the drop in the group average similarity of both candidate groups, with the given threshold thr_{drop}*:*

The current check point analyses the prospective likeness structure of the larger merging group C that is composed of both groups A and B. Hence, the drop in group average similarity value of both candidate groups $Group_A = \{seg_m, seg_d, seg_e\}$ and $Group_B = \{seg_n, seg_a, seg_b, seg_c\}$, can be expressed as:

$$Drop_A = groupAveSim_C - groupAveSim_A \tag{6.22}$$

$$Drop_B = groupAveSim_C - groupAveSim_B \tag{6.23}$$

where, $C = \{seg_a, seg_b, seg_c, seg_d, seg_e, seg_m, seg_n\}$. If both calculated drops are below a given threshold thr_{drop}, i.e., $Drop\ A < thr_{drop}$ and $Drop\ B < thr_{drop}$, then the merging criterion is reached and the entire segments of both candidate groups A and B are merged into a larger cluster C. Otherwise, the algorithm does not allow the merging between these two candidate groups, followed by the proceeding of a new clustering iteration to find the next highest similarity and the next closest admission pair of segments.

In the next step, as the joining procedure is accomplished, the group average similarity for the newly formed cluster C is updated as $groupAveSim_c$. Then, as it can be seen in Figure 6.19 and 6.20, the new mutual overlap paths for each segment of cluster C can be calculated as follows:

$$overlap_{seg_i} = overlap_{seg_i}\ \cap\ [max(st_{seg_i}) : min(end_{seg_i})] \tag{6.24}$$

where, $i = \{a, b, c, d, e, m, n\}$, st_{seg_i} and end_{seg_i} are set a of starting and ending points that were obtained at the previous check point.

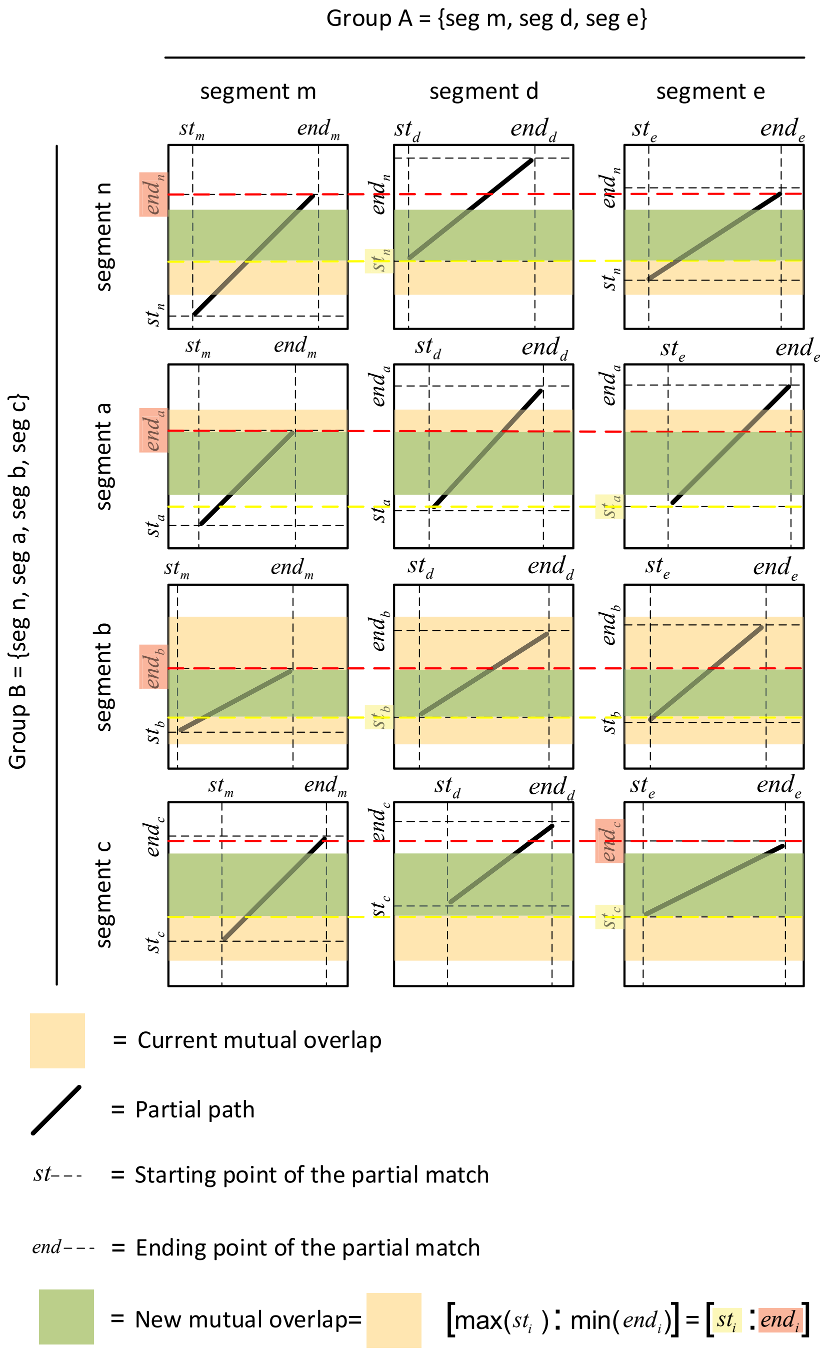


Figure 6.19: Updating mutual overlap path for the segments that were associated to the previous group B.

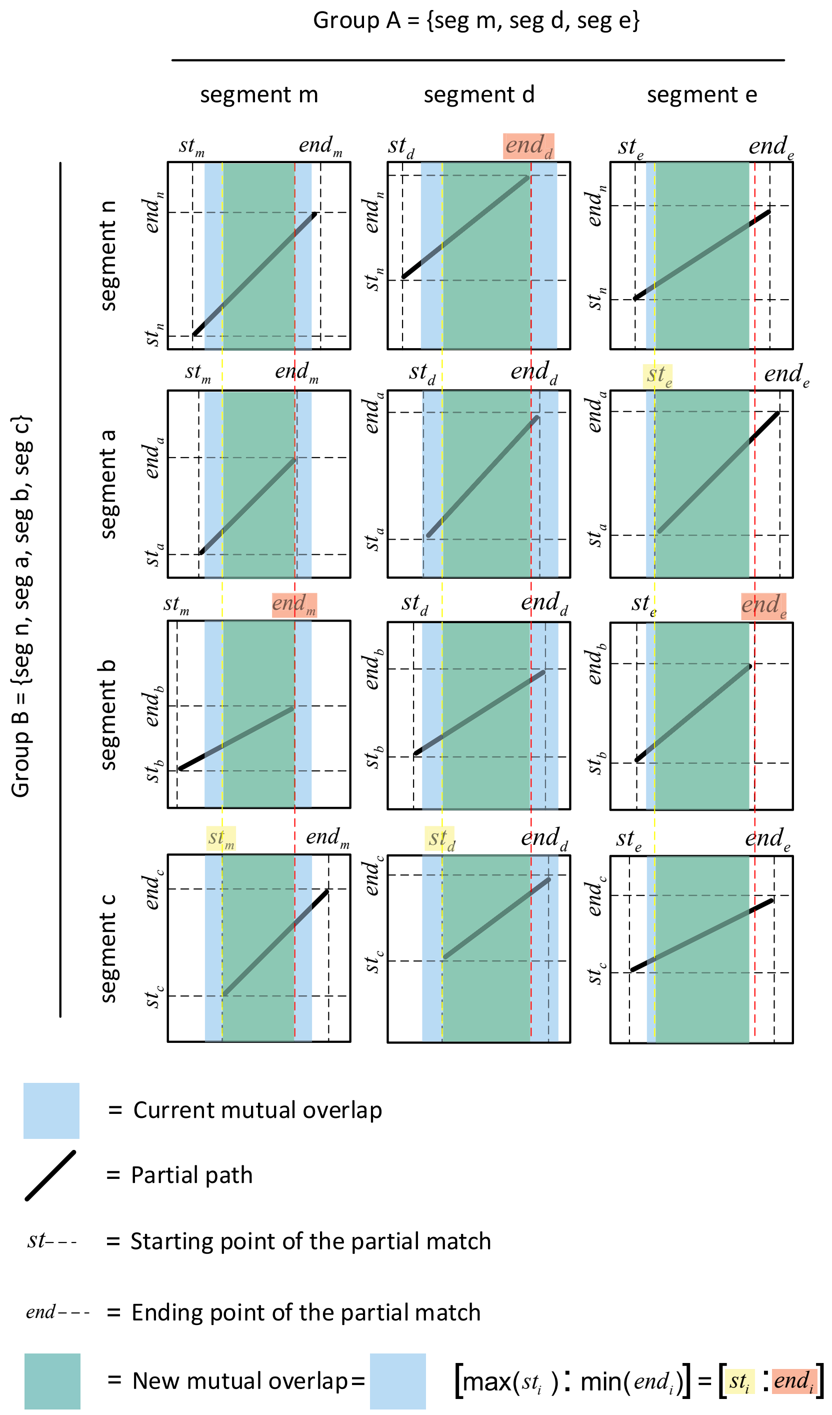


Figure 6.20: Updating mutual overlap path for the segments that were associated to previous group A.

3- Termination criterion

The procedure of merging into larger clusters and correspondingly updating the clustering relationship information is repeated recursively until the pairwise similarity of the next closest pair of segments, is not available or is below a pre-defined value, i.e. $simDisMat(m, n) < D_{thr}$. In the latter, the remaining single segments, which is not assigned to any cluster with several segments, are assumed as a distinct single-member group.

6.3.3 Clustering output results

The outcome of the modified unsupervised clustering at the reached termination level can be expressed as a set of distinct clusters $C = C_1, \cdots, C_n$, and the entire data instances can be formed as $S = \bigcup_{i=1}^{n} C_i$, where $C_i \cap C_j = \varnothing$ for $i \neq j$.

Each obtained clusters is described with a series of indices of its own members followed by evaluating a corresponding representative pattern. The representative pattern is a time series vector and it is used to illustrate the overall shape of its corresponding cluster. Moreover, this representative pattern vector or *repPat* is employed as a new input segment for further DTW searching within different sound files of a particular bird species in the data and it will be explained in the following experimental evaluations' section.

The representative pattern vector of each cluster can be evaluated from the following two steps:

Step 1: finding the two closest segments in the cluster which have the largest pairwise similarity within the entire members. For simplicity purposes, consider that cluster $C = \{seg_1, seg_2, \cdots, seg_n\}$ and seg_1 and seg_2 are the two closest segments where, $max(sim_i) = sim(seg_1, seg_2)$, for all $i = 1, \cdots, n$.

Step 2: the representative pattern vector is selected partially from the first segment of the closest pair or seg_1 as: $repPat = seg_1(st_1 : end_1)$, where st_1 and end_1 are the starting and ending frame of seg_1 and they can be defined from the partial path matrix as,

$$parPathMat(seg_1, seg_2) = \begin{bmatrix} st_1 \\ end_1 \\ st_2 \\ end_2 \end{bmatrix} \tag{6.25}$$

6.4 Experimental evaluations

This section describes the experimental evaluation of the entire system that incorporate both the described methods; the modified DTW similarity calculation and the hierarchal unsupervised clustering method to discover a set of structurally distinct acoustic units of each particular bird's vocalisations dataset.

6.4.1 Data description

The outcome of the sinusoidal detection method (as described in Chapter 5), provides a set of distinct frequency tracks segments within each recording file. Thus, these obtained segments, within each particular bird species dataset, are used as a means of temporal input sequences for the entire proposed vocalisations' discovery system. Moreover, since there is no element-level label information available with the data, the result of this evaluation system then will be used for training purposes in our bird vocalisations recognition system, as described in Chapter 7. At the beginning of the following experiments, the whole vocalisations' data is split into training and testing parts. The selected partition ratio of training to testing is 2 : 1. In this implementation procedure, every three seconds of data is split into three different folds of frames, of which the first two folds and the third fold are assigned to training and testing data, respectively. Hence, in order to have labelling information for building the recognition system, the following experimental evaluations are only performed on frequency tracks segments of training data.

6.4.2 Procedure of experimental evaluations

In order to perform the experimental evaluations within each particular bird species, the whole procedure in each bird subset, initially proceeds in every particular recording file; then it continues within the entire files of each bird species. As it can be seen in Figure 6.21, at the first stage, the frequency tracks segments are used as a means of temporal input sequences for both DTW similarity calculating and clustering purposes. The outcome of the modified clustering method provides a set of information for every obtained cluster, including the representative pattern vector with corresponding information about the group occupancy and the list of associated segments. All the obtained pattern vectors with the other remaining single segments in each separate file are then used as a new set of individual segments. The procedure in the second stage is the same as the previous stage; while, in the final outcome of the system, instead of keeping the pattern vectors' indices, the indices of all assigned segments which are linked with each corresponding pattern vector, are replaced by using the clustering information of stage one (clustering information result that was obtained in each recording file).

In general, the outcome of the second clustering function provides some useful information about every obtained cluster over each particular dataset (bird species); such as: list of segments in each group, list of single segments and other relevant group occupancy within each level of clustering.

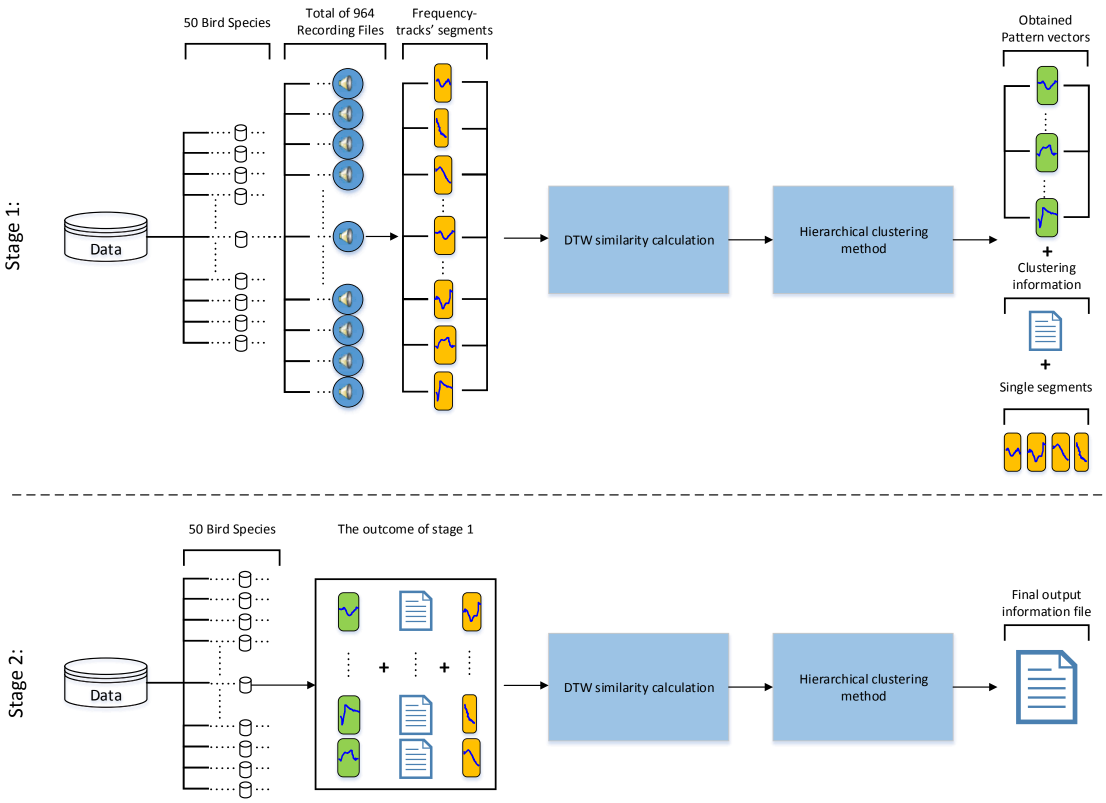


Figure 6.21: Diagram of the procedure of experimental evaluations

Parameters' setup

In the modified DTW algorithm, first, the value of the local constraint's parameter C was set to 2 for vertical and horizontal movements on the warping path. In the partial DTW searching procedure, the value of D_{thr} and L_{min} (minimum length) were set to 1.2 and 15, respectively. Moreover, the value of j_{step} was set to 4 frames in each continuous and new

DTW searching iteration. In the similarity calculating procedure, the steepness and mid-point's coefficients α_D and β_D of $P(x_D)$, were set to -3 and 1.2, respectively empirically. In the next probability function $P(x_L)$, coefficients α_L and β_L can be estimated based on the overall lengths in the detected partial path of each particular recording file. Hence, these values are calculated automatically in the DTW searching procedure separately, for each given set of segments as follows:

$$\beta_L = mean(\sum_{i=1}^{n} Length_{parPath_{(i)}}) \tag{6.26}$$

$$\sigma = std(\sum_{i=1}^{n} Length_{parPath_{(i)}}) \tag{6.27}$$

$$a_L = \frac{log(\frac{0.95}{1-0.95})}{2 \times \sigma} \tag{6.28}$$

where, *mean* and *std* are the mean and standard deviation function, respectively; $Length_{parPath_{(i)}}$ represents the corresponding length of the i^{th} partial path; n is the total number of the detected partial paths in each recording file; and σ is the standard deviations value.

In the clustering procedure, the values of thr_{drop} and D_{thr} were set to 0.13 and 0.5, respectively, as they were found empirically.

6.4.3 Experimental results

As, the data which is used in this research doesn't have any element-level labelling information (time-stamps), therefore, the entire unsupervised discovery system can not be evaluated numerically. Furthermore, the effect of clustering procedures is evaluated in terms of bird species recognition accuracy in Chapter 7.

In general, the obtained result of the whole proposed bird vocalisation learning system could not be investigated or analysed separately before any incorporation among the bird recognition systems, where this clustering information is considered as further labelling information. Hence, the corresponding outcome of the proposed system can be observed along with the possible improvements on further recognition accuracy.

In the following section, the clustering grouping results in terms of occupancy attracting rate on the whole data and each bird species separately are observed.

Results in stage 1:

As it mentioned in 6.4.2, in the first stage of the experimental procedure, the proposed system was performed on the entire segments of each individual recording file. By employing the DTW searching outcomes such as, the DTW partial path's information and the similarity distance matrix, the proposed clustering method provides a set of district clusters for each recording file. Then, each initial cluster can be presented with the list of its own corresponding segments.

The obtained results in clusters and the other remaining segments will be used in further experimental evaluations; thus these results are considered as an initial outcome. Also, as these evaluations were performed on a large data set with more than 900 recording files, for the purpose of visual inspection of the output result, only the result of three recording files are observed at this stage and the final occupancy result within each sub data set will be discussed more precisely in the next stage.

Three candidate files with file ID's, 28321, 4588 and 1396 in our data, are selected from the data set number 10 with the corresponding labelling name of,"Nashville Warbler" (more details are available in appendix A and B). The initial clustering results of these recording files are discussed as follows.

Clustering outcome of recording file 28321:

The first recording file is composed of 55 different segments, 33 of which belong to the training data. The obtained frequency tracks and corresponding clusters are depicted in Figure 6.22. At this stage, the segments of this file are grouped into five multi-member clusters as $cluster_1, \cdots, cluster_5$; where the assigning occupancy in each cluster is between 6 and 7 and there is only one segment remaining as a single cluster. As it is demonstrated in Figure 6.22, all the individual segments in each cluster have a homogenous vocalisation pattern. Hence, it can be assumed that a set of structurally distinct bird vocalisation patterns are obtained in this single recording file. Investigation on both length and frequency pattern of the remaining single segment shows that this segment is not a bird vocalisation and this segment can be considered as a vocalisation noise.

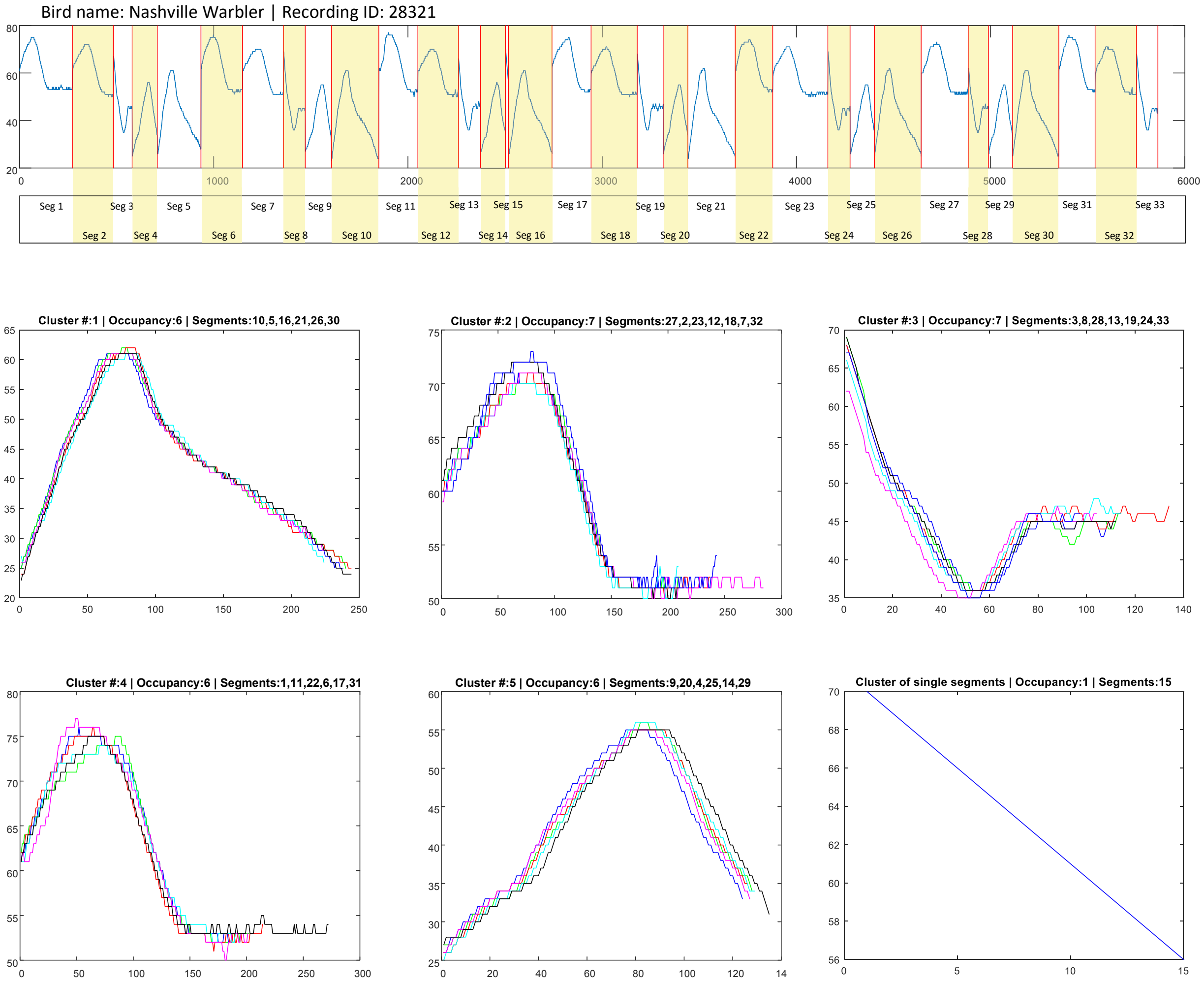


Figure 6.22: The clustering outcome of file 28321 from data set number 10, with illustration of frequency tracks (on top) and the obtained clusters.

Clustering outcome of recording file 4588: The second recording file is composed of 83 different segments, 60 of which belong to the training data. The obtained frequency tracks and corresponding clusters are depicted in Figure 6.23. At this stage, the segments are grouped into ten multi-member vocalisation clusters as $cluster_1, \cdots, cluster_{10}$; where the occupancy in each cluster varies between 2 and 13 and the cluster of single segments consists of 10 single segments. About 40% of the segments are in the first two clusters with the occupancy above 12 ($cluster_1$ and $cluster_2$) and only less than 17% are remain as a single segment. Further investigation on the single segments and obtained clusters shows that the employed joining rules in the algorithm did not allow the clusters to attract these single segments, as their corresponding length or frequency range does not match to the singles. There is only one miss-matching segment found (segment 28 with the yellow pattern), as the corresponding pairwise similarity of this single segment among the other individuals in $cluster_4$ is less than the termination threshold. As it can be seen in Figure 6.23, a set of structurally distinct bird vocalisation patterns are obtained in this single recording file.

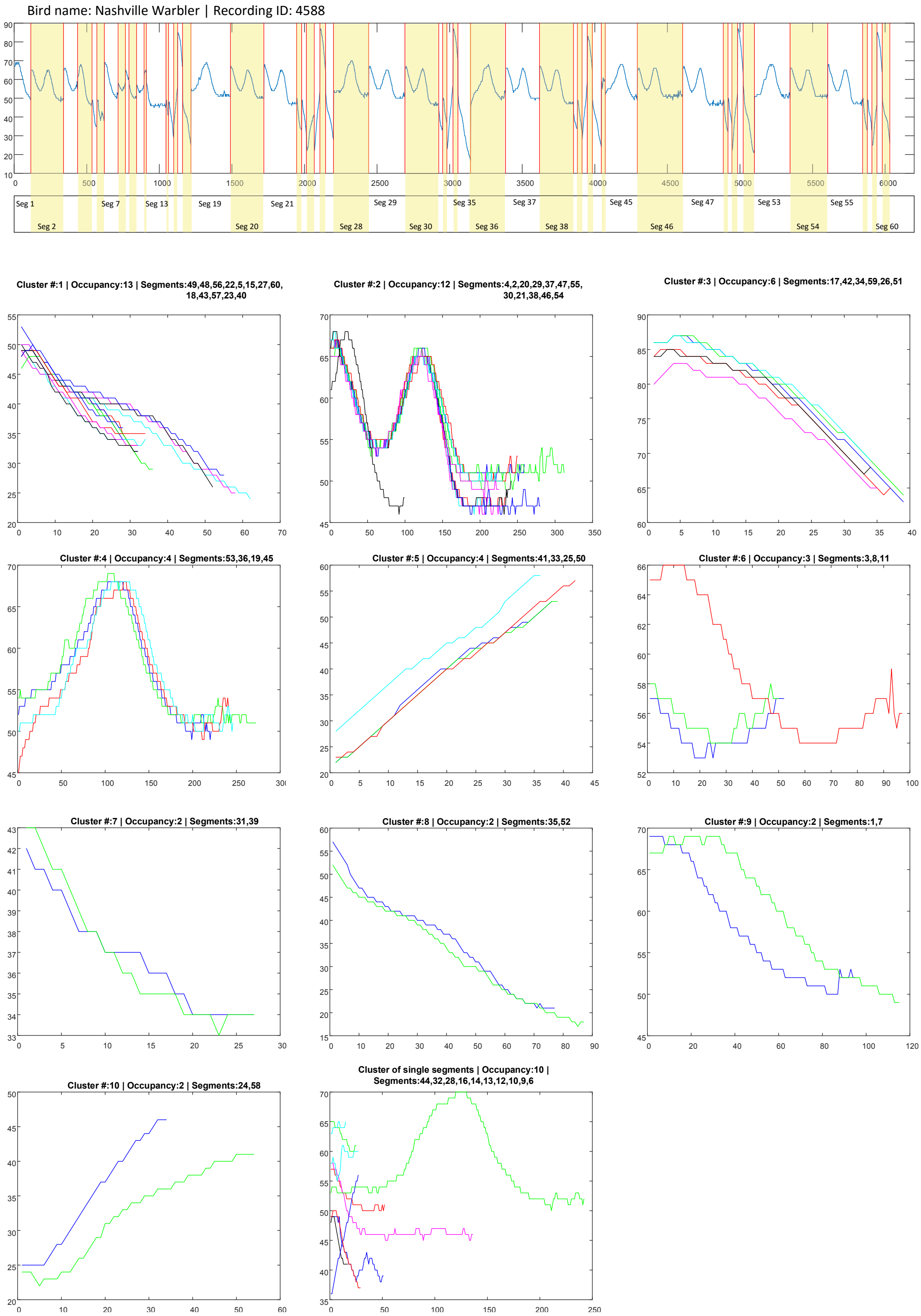


Figure 6.23: The clustering outcome of file 4588 from data set number 10, with illustration of frequency tracks (on top) and the obtained clusters.

Clustering outcome of recording file 1396: The third recording file is composed of 34 different segments, of which 21 belong to the training data. As it can be seen in Figure 6.24, the initial resulting clustering shows that the segments of this short file are grouped into three multi-member clusters: $cluster_1$, $cluster_2$ and $cluster_3$. More than 50% of the segments occupied $cluster_1$, all with a similar frequency range and pattern. The other two clusters each have three segments. Also, the cluster of single segments consists of four segments. By analysing the length and the frequency range of each single segment, it is considered that the segments may correspond to a variety of other acoustic events such as tonal noise and other bird/animal vocalisations in the background.

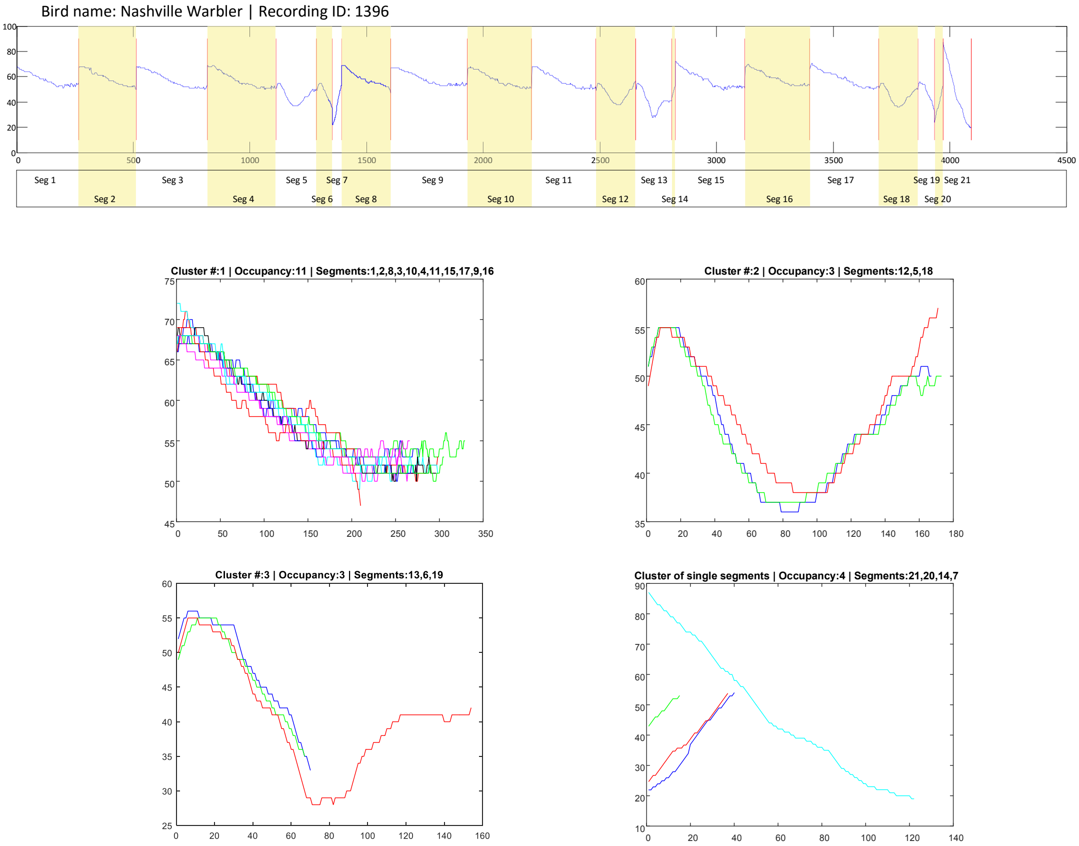


Figure 6.24: The clustering outcome of file 1396 from data set number 10, with illustration of frequency tracks (on top) and the obtained clusters.

In general, the investigation on the above recording files shows the good coherence of the frequency range and the pattern structure in each cluster.

Results in stage 2:

1-Visual inspections: A part of the obtained clustering result for data set number 10 (bird 10) is illustrated in Figure 6.25. In the figure, each row corresponds to an individual cluster found and each column shows an example of the frequency track segment that is associated with that cluster. The rows are sorted in terms of the occupancy; hence the first row represents the largest found cluster in the collection data 10. As it can be seen from Figure 6.25, frequency tracks within each cluster show great similarity to each other, while across clusters show clearly distinctive patterns.

2-Occupancy inspection: The occupancy information of the final obtained vocalisation clusters, when using the above parameters' setup, for each data set can be illustrated in Figures 6.26, 6.27 and 6.28. Each graph in the figures displays the occupancy of the hundred (or less in some bird species, e.g. 32 and 38) largest clusters, in decreasing order. Meanwhile, the clusters are presented with the indices varying between 1 and 100, as the largest and smallest cluster, respectively. In addition, the percentage value in each section (between two red lines), shows the occupancy proportion rate of each 10 clusters to the total number of the training segments in each data set. For instance, as it can be seen in Figure 6.26, plot 1, there are 10 clusters with an occupancy above 150 segments, which means that over 19.3% of the detected segments in dataset 1 occupied this set of clusters. Also, over 50% of the training segments of this bird species were assigned to the first 40 largest clusters with an occupancy above 70 segments.

By investigation of the attracting ratio for the first 10 clusters of each data set, it is

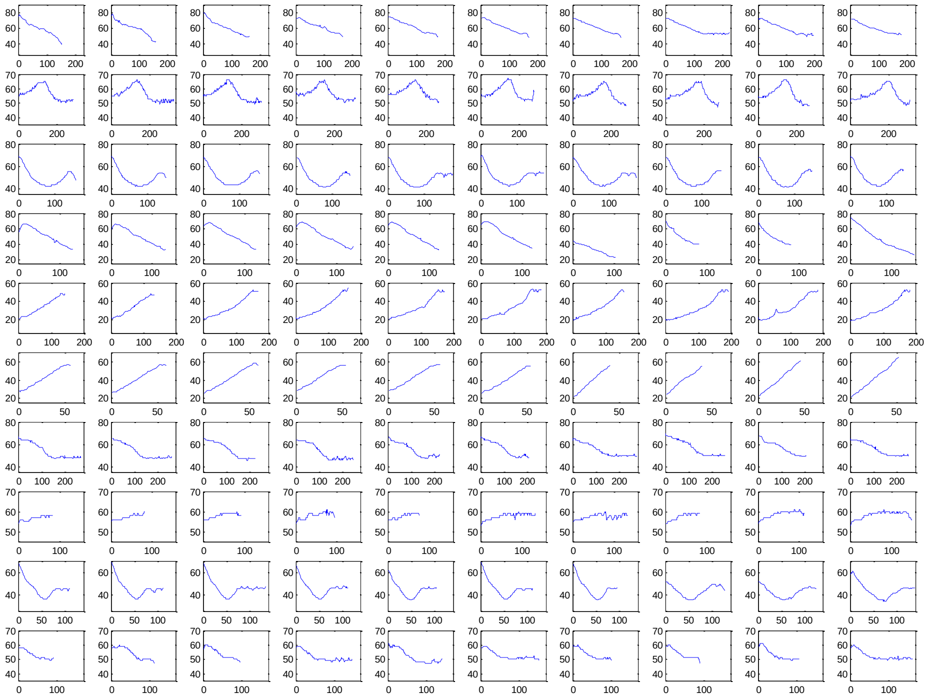


Figure 6.25: A part of the outcome of the unsupervised clustering depicting several examples of frequency tracks (where the x and y-axis corresponds to the frame-time and frequency index, respectively) of segments associated with ten different clusters (corresponding to each row).

considered that bird 32 has the best attracting ratio, as more than 75% of the detected segments are assigned into the first 10 largest clusters; whereas, bird 22 has the lowest attracting ratio with 8.47%. Additionally, the first cluster in bird 12 has the highest occupancy in the whole data.

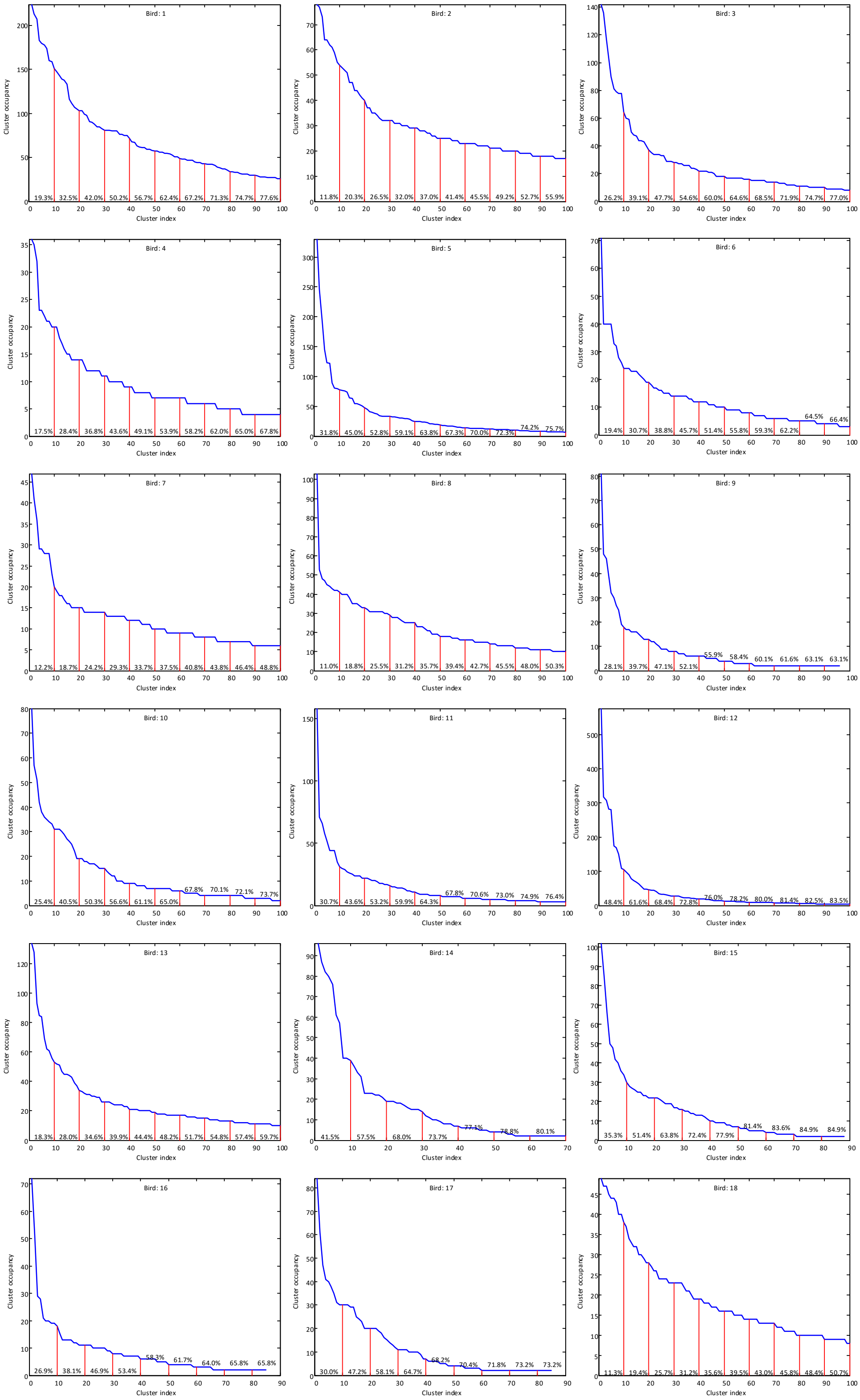


Figure 6.26: The relative occupancy information for data set 1 to 18

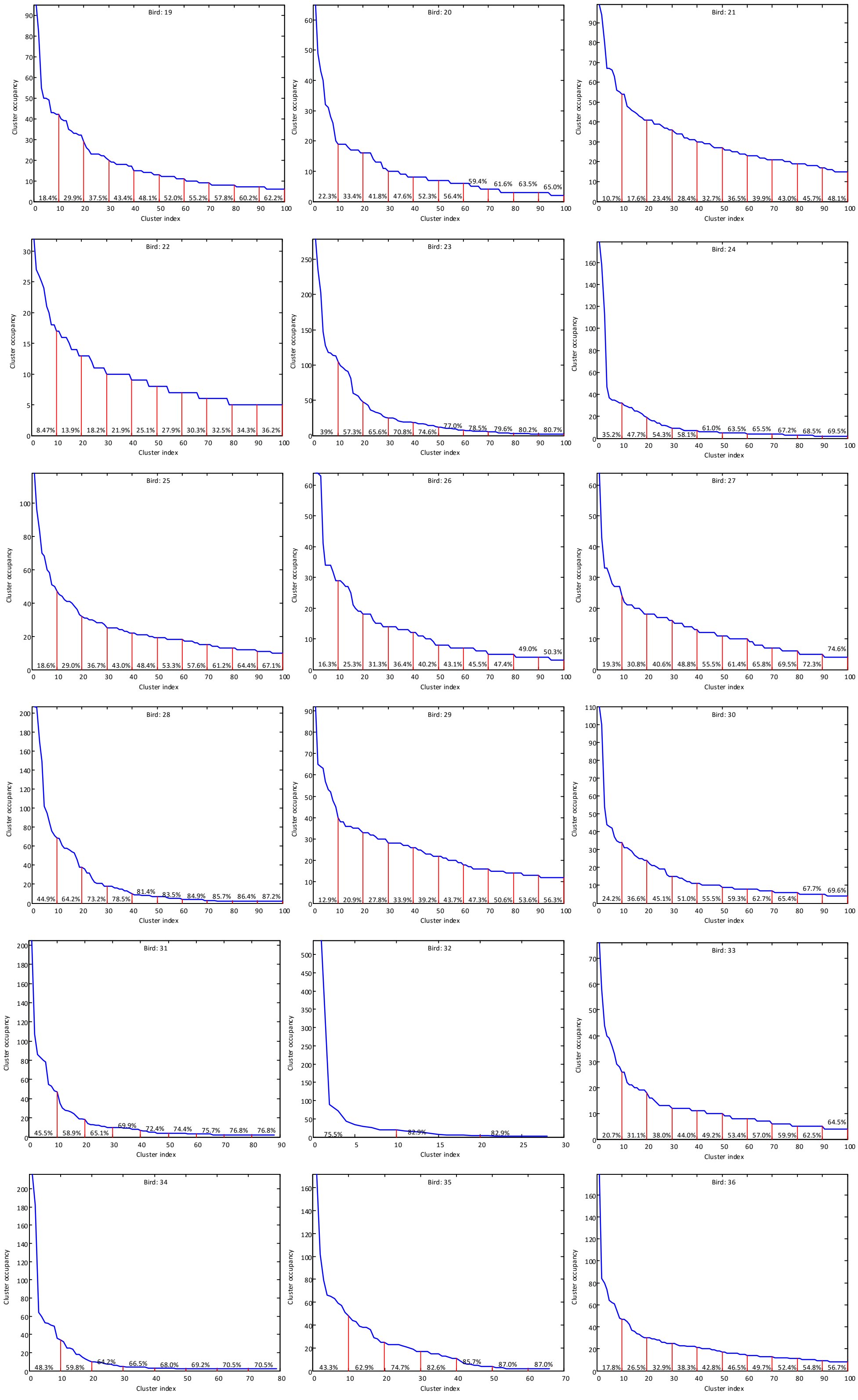


Figure 6.27: The relative occupancy information for data set 19 to 36

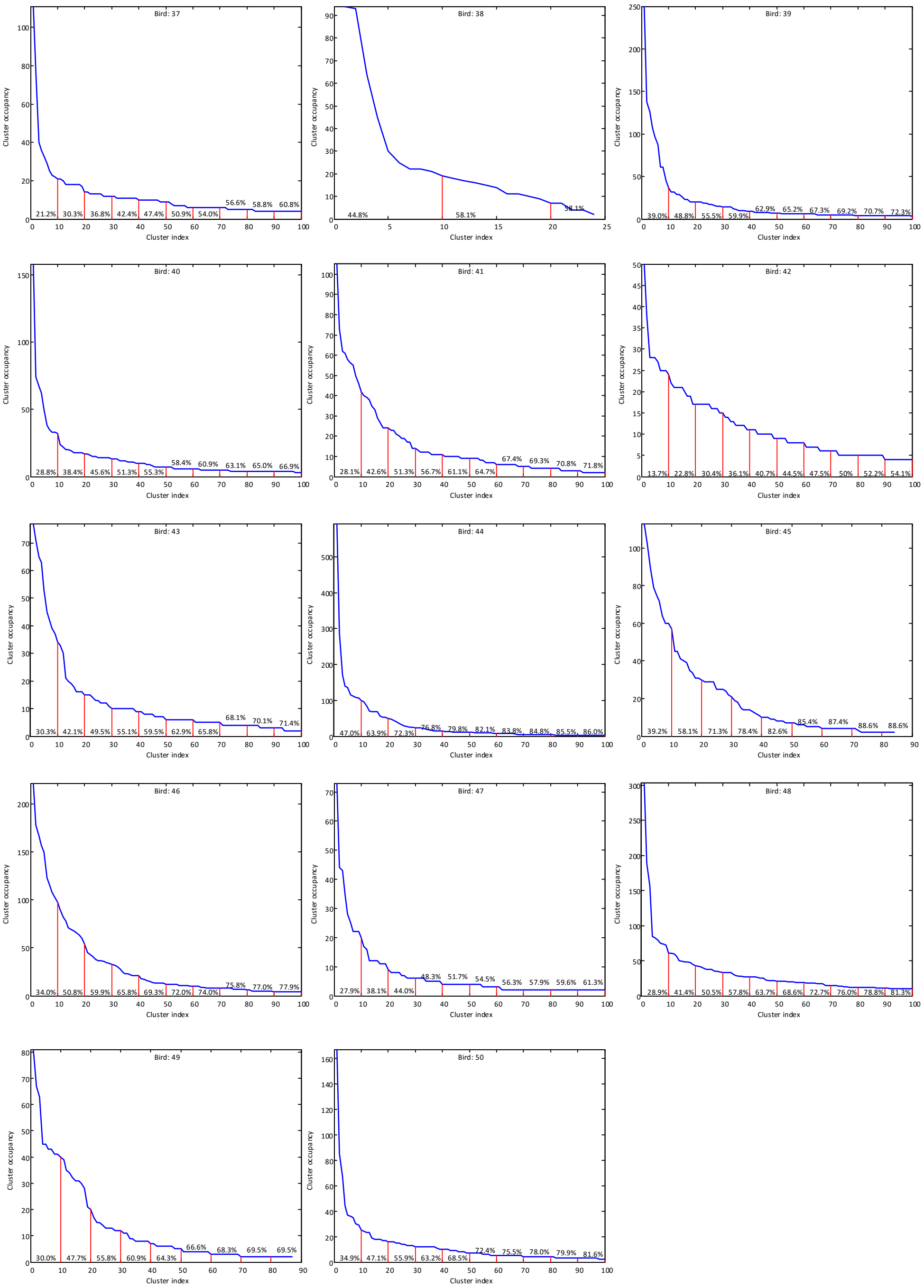


Figure 6.28: The relative occupancy information for data set 37 to 50

The occupancy information of all 60 largest clusters of each bird species along with the size information of each data collection can be summarised in Figure 6.29. As can be seen in this figure, the black bars indicate the total number of segments in each data collection, while the coloured bars show the total number of segments within each corresponding occupancy division. For instance, bird 12 is composed of more than 5100 segments and about 3990 and 2400 of these segments are attracted into the first 60 and 10 largest clusters, respectively.

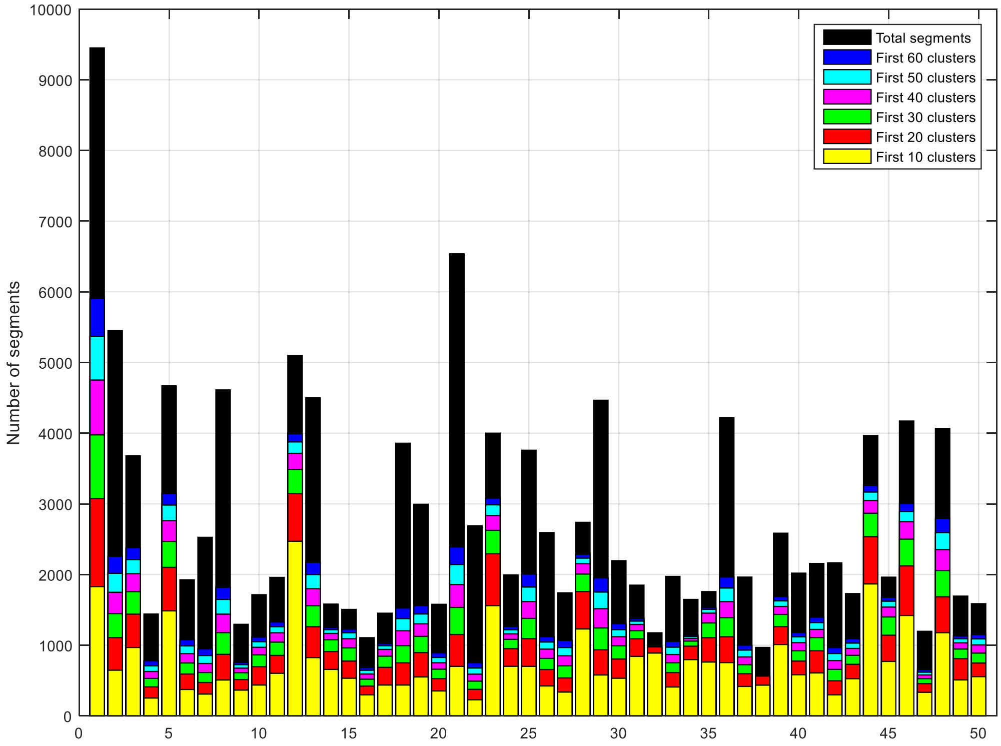


Figure 6.29: Illustration of the total number of training segments for each defined clustering proportion of each data set 1 to 50

The overall clustering occupancy of the first 10 to 80 largest clusters among the whole corpus data is presented in Figure 6.30, by obtaining the average and the standard deviation (above and below) of the entire corresponding occupancy values over all the bird species. As it can be seen in Figure 6.30, about 30% and 70% of the detected segments of the whole training data are assigned into the first 10 and 80 largest clusters, respectively.

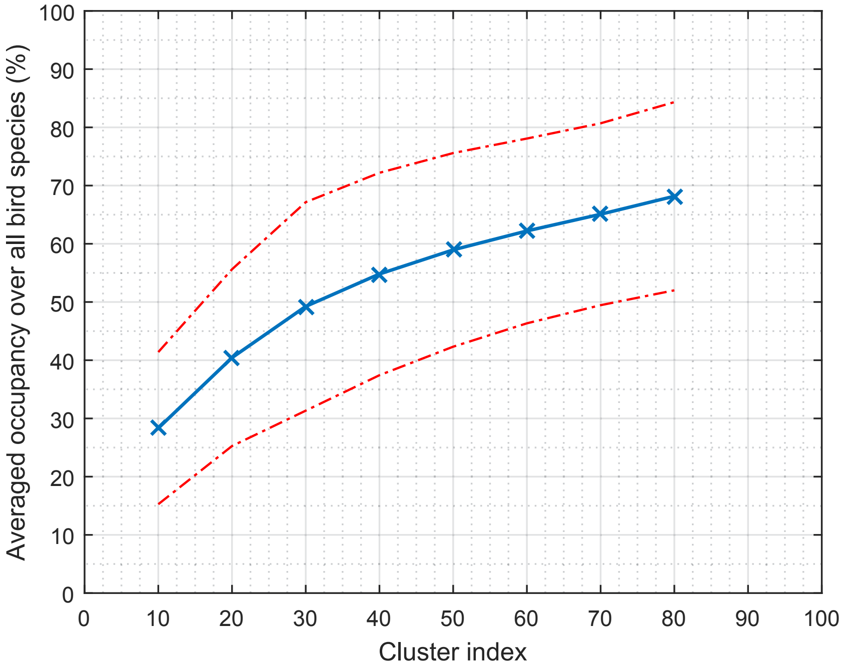


Figure 6.30: Illustration of overall clustering performance of first 10 to 80 largest clusters within the entire corpus data

6.5 Summary

In this chapter, an approach for unsupervised discovery of acoustic patterns in bird vocalisations is proposed. This approach employed the detected frequency tracks in Chapter 5, as features to characterise bird tonal vocalisations. The whole discovery system was composed of two consecutive parts. First, a modified DTW algorithm that enabled the search for multiple and partial matchings between two segments was developed. For each pair of sequences, the modified DTW algorithm was then employed in several parallel searches, each performed from a different starting point on one of the sequences and allowed to start anywhere on the other sequence. Then, the similarity score for each obtained partial matching path was calculated based on a combination of the normalised cumulative distance and the length of the DTW matching path. The similarity calculation was completed by keeping out of these paths only a match with the highest similarity score. This selected path with its associated similarity score was then considered as a final partial DTW searching outcome, for each given pair of segments. After processing the partial DTW searching of all the detected segments, the DTW searching outputs with the entire detected segments (of each particular bird species) were then used in a novel proposed hierarchical clustering approach, to group all the homogeneous structured segments into a set of distinct element-based vocalisation clusters. Several joining rules and conditions were employed in the proposed clustering method to control the joining procedures. The merge decisions were always based on further investigations on the prospective likeness structure of the group, including both the merging objects. Finally, experimental evaluations demonstrated that the obtained clusters showed good coherence and provided a set of structurally distinct bird vocalisation patterns.

Chapter 7

AN AUTOMATIC HMM-BASED BIRD SOUND RECOGNITION SYSTEM

7.1 Introduction

This chapter presents two automatic HMM based recognition approaches as baseline and element-based systems, for identification of bird species from the natural field recordings, by using the detected vocalisation elements (frequency tracks segments) as temporal sequences.

7.2 Building an automatic bird species recognition system

The feature extraction process along with the segmentation procedure, as explained in Chapter 5, deliver a set of detected segments for a given audio file of the corpus data. As these obtained frequency tracks represent the vocalisation's element units (the smallest structurally stereotyped acoustic units produced by bird), they are considered as element segments in this research.

For the purpose of building a bird species recognition system, HMM is employed to provide the model(s) for each bird species. Each corresponding model is built based on modelling the temporal evolution of frequency tracks of elements, by employing a left-to-right topology in each HMM. Moreover, Gaussian distribution(s) with a diagonal covariance matrix, which are used widely in audio/speech pattern processing, are preferred as the HMM state output probability density functions (pdf) for computational purposes [134].

In the following sections, two HMM recognition systems are presented; the base line system where a single HMM model is provided for each bird species, and the proposed element-based system, where a set of single HMM models is provided for individual vocalisation elements for each bird species.

7.2.1 Baseline HMM-based system

In the HMM baseline system, the entire detected segments within the training data set of each bird species, are modelled with the single HMM model. The probability density function in each state is modelled by mixture of Gaussians, to allow for the collection of element pitch patterns and the diversity of individual entities of vocalisations [134].

7.2.2 Individual element-based HMM

In this section, a element-based HMM recognition system is presented. In the proposed system, instead of employing a single model for each bird species, the aim is to build a separate HMM to model each type of bird vocalisation element for each bird species.

Element-level label information

Obtaining the individual element models is straightforward if the labelling metadata for the elements was available or if the set of vocalisation patterns produced by each bird species was known [134]. Accordingly, there is no such information available in our corpus data (it is rare for it to be available for other large corpus typically). As such, a question is raised here, of how the individual element HMMs could be trained in the aspect of an unsupervised procedure? The described unsupervised discovery of elements in the bird vocalisation method in Chapter 6 deals with this problem, by providing a set of clusters of element patterns for each bird species. Therefore, the resulting outcome also offers the label information for each discovered segment of the data [134].

Modelling individual bird elements

By employing the above mentioned element-level information, a fixed number of clusters, based on the highest occupancy, are used for the purpose of training the individual element HMMs of each bird species. As it is assumed that each obtained bird vocalization cluster is composed of a set of similarly structured patterns, only a single Gaussian distribution is used as the state output probability density function (pdf) of each individual element HMM.

Moreover, to the above individual element HMMs, an additional single HMM model is employed to model all the remaining segments in clusters which are not assigned into those selected largest clusters, along with all the remaining single segments in each bird species. For fitting the variety of these remaining segments, several Gaussian mixture components are used as the state output pdf of this HMM [134].

The Baum-Welsh algorithm is employed to train all the mentioned individual element HMMs. Furthermore, Figure 7.1 illustrates the state output pdf of nine trained individual element HMMs of two particular bird species.

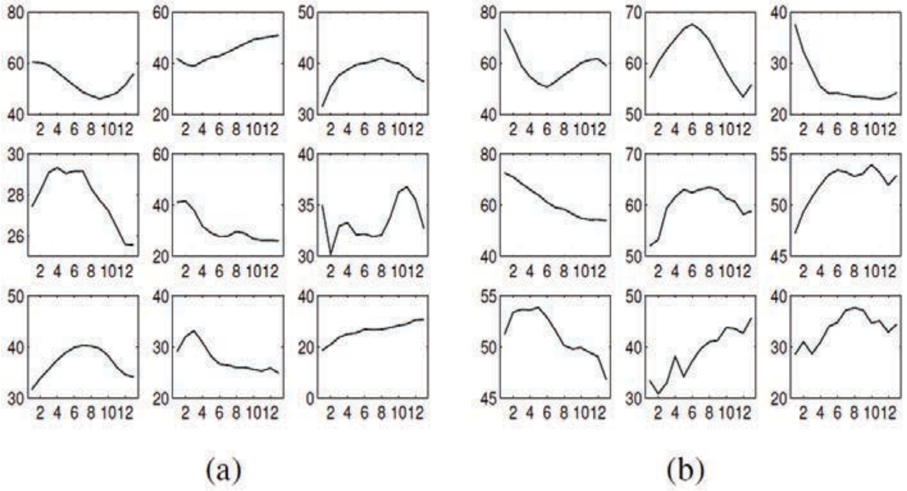


Figure 7.1: Illustration of of the mean values of the state output Gaussian pdf, modelling frequency track features, for nine trained element HMMs of bird species House Finch (a) and Northern Cardinal (b). The x- and y-axes denote the HMM state and frequency index, respectively [134].

7.2.3 Recognition of bird species

The goal of the recognition step, in both presented systems, is to identify the bird species from a finite set of species based on a given utterance of a test signal.

The recognition stage starts with providing a set of N detected frequency tracks segments for each utterance with a given length. This set of test segments are obtained with the segmentation and feature extraction step (as described in Chapter 5) and can be expressed as, $O = \{O_s\}_{(s=1)}^{N}$, where each segment s is a sequence of features $O_s = \{o_s^1, \cdots, o_s^{T_s}\}$, where T_s defines the number of frames in segment s. The recognition procedure is treated based on each detected segment, separately, as the Viterbi method is employed to obtain the probabilities of each corresponding test segment s on each bird species model λ_b as, $p(O_s|\lambda_b)$ [134].

The recognition procedure in the proposed element-based HMM system is progressed by calculating the probability of O_s on each individual element model of each bird species, followed by taking the maximum.

By assuming that all the vocalisations in the given test utterance O_s belong to only one single bird species, the probability of the utterance being performed by each bird species b, can be calculated as the product of the individual segment probabilities as, $p(O|\lambda_b) = \prod_{s=1}^{N} \; p(O_s|\lambda_b)$ and the identification can be obtained by [134, 53]:

$$b^* = arg\ max_b\ p(O|\lambda_b) \tag{7.1}$$

For the assumption that there is a possible existence of other birds/animals' vocalisations in the given test utterance O_s, other than the current species in the vocabulary list, the further calculation of the overall probability $p(O|\lambda_b)$ has been done in a similar way to [53, 138]; by removing the product of those segments, where their probability was below a given threshold in all the models. The observed results show no improvement in the system.

7.3 Experimental evaluations

7.3.1 Data description

Experimental evaluations were performed on all the outcomes of the frequency tracks detection and segmentation step presented in Chapter 5. Also, the outcome of partial DTW searching and the modified hierarchical clustering method are used in the following experiments as corresponding labelling metadata for the detected elements.

For the purpose of training and testing, as mentioned in Chapter 6, the entire bird vocalisations' corpus data is split into two parts and the selected partition ratio of training to testing is 2 : 1. In other words, every three seconds of each recording file is split into three different folds of frames, of which the first two folds and the third fold are assigned to training and testing data, respectively. Moreover, the testing data is divided further into a set of utterances, where each utterance is composed of a signal containing approximately the given length of the detected segments [134]. Figure 7.2, illustrates the summary of data which are used in the proposed recognition systems.

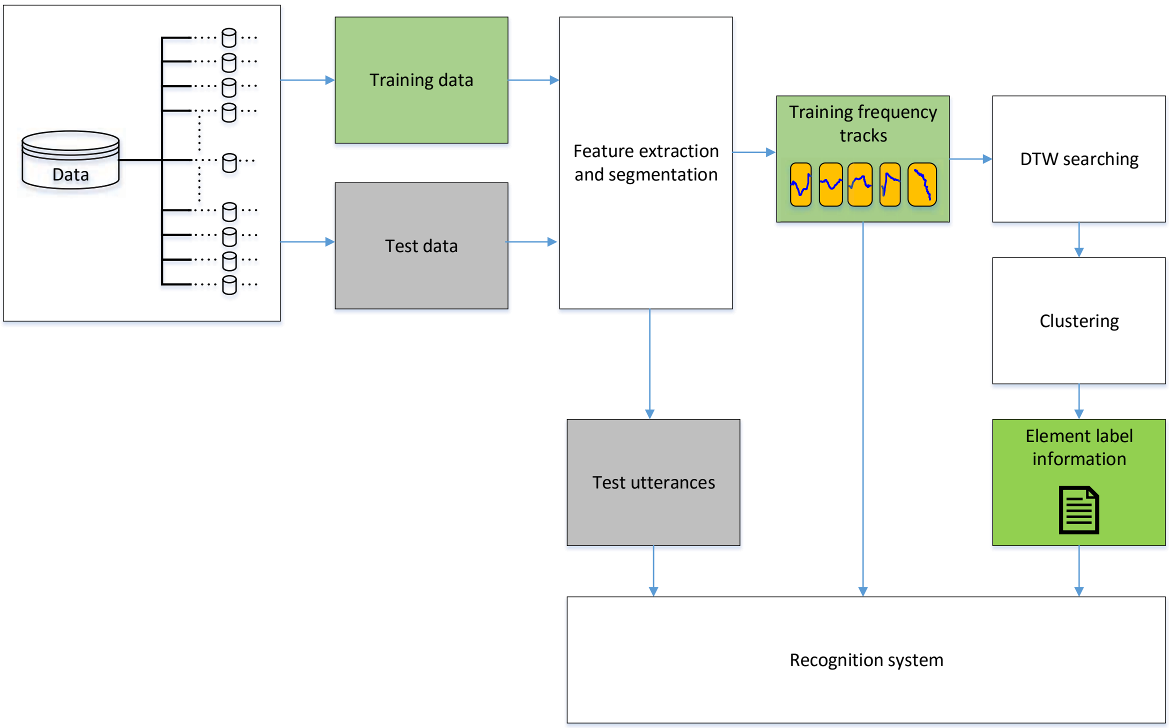

Data
Training data
Test data
Feature extraction and segmentation
Training frequency tracks
DTW searching
Clustering
Element label information
Test utterances
Recognition system

Figure 7.2: The data description used in the proposed recognition system

As the minimum required number of obtained clusters for each bird species is set at 40 in further experiments, a set of 48 (out of 50) bird species are selected from the above corpus data. These are selected in such a way that for each bird species at least a set of

40 bird vocalization clusters are available. As it can be seen in Figure 6.26 (in Chapter 6), the entire segments of two bird species number 32 and 38, are grouped into less than 30 clusters. As a result, these two subsets are skipped for the further experiments.

7.3.2 Experimental setup

Each frame time of the detected frequency tracks sequences in Chapter 5 only shows the obtained frequency information (see section 5.1). As the current frequency features do not include any information to describe how the features derive over the time index, and in order to add local dynamic information such as the temporal time derivatives to the statics parameters, the delta and acceleration features are calculated in [139] with the window parameters set to 3 and 2, respectively. The resulting derivatives are then added to the corresponding frequency tracks' statics, in such a way to form 3-dimensional feature vectors. Furthermore, in the following experiments, the number of states in each HMM model is set to 13, in order to reflect the minimum allowed length of the detected elements [134].

7.3.3 Experimental results of baseline recognition system

As it can be seen in Table 7.1 the result of the HMM baseline system is achieved by the number of Gaussian mixture components at each HMM state. A different varying utterance length of 1, 2 and 3 seconds, is observed in the experimental evaluations.

As Table 7.1 shows, the recognition accuracy is quite high for 10 mixture components. The resulting recognition accuracy gradually increments as the number of mixture components increase up to 60 and then become flat.

Table 7.1: Bird species recognition accuracy (RA) achieved by the baseline HMM-based system using a single model for each bird species with a given number of mixture components per state.

Utterance length (sec)	Number of mixture components per state										
	10	20	30	40	50	60	70	80	90	100	110
	Recognition accuracy (%)										
1	68.75	72.30	74.36	75.21	75.67	76.02	76.03	75.95	76.00	76.22	76.38
2	75.89	79.91	81.40	81.60	82.03	82.19	82.31	82.18	81.85	82.02	81.99
3	80.30	83.97	84.66	85.12	84.92	85.25	85.27	85.40	85.48	85.73	85.55

7.3.4 Experimental results of element-based recognition system

In this section the achieved results with the proposed element-based recognition system are presented. As mentioned previously, the state output pdf for each individual element HMM was composed only of a single Gaussian distribution; whereas, the additional model that is used to cover all the remaining segments in each bird species, employed a mixture of Gaussians with m number of components. In order to compare the new models to the baseline model, as presented in Table 7.1 at the first step a fixed number of individual models as parameter n, are used for all bird species. For each experiment, the values of both n and m were set in such a way that the summation of these two parameters is equal to 60, in order to have comparable conditions to the baseline model with 60 Gaussian mixture components per state used [134]. Using different values for the corresponding number of individual models (n) and number of mixture components (m) gives the outcome results in Table 7.2.

Table 7.2: Bird species recognition accuracy (RA) obtained by the element-based HMM system using individual models of bird elements.

Utterance length (sec)	Baseline system	Element-based HMM system				
		Number of individual element models				
		10	20	30	40	50
		Number of mixture components per state for single HMM				
		50	40	30	20	10
	Recognition accuracy (%)					
1	76.02	79.38	82.12	84.85	86.33	86.67
2	82.19	84.94	87.43	90.10	91.01	90.54
3	85.25	87.93	90.13	92.52	93.44	92.90

As can be seen in Table 7.2, the resulting recognition accuracy with the proposed element-based system improved considerably; while the complexity of the models of both systems is identical. Furthermore, the recognition accuracy increases gradually while the number of used individual element models increments up to 40 and then becomes flat.

Another similar set of experiments were performed to analyse the proposed system when the given number of individual models (n) is not fixed within all bird species. Therefore, each value of parameter n for a particular bird species was calculated based on the corresponding occupancy of each cluster, by defining the threshold in a way that the number of individual models is 40 on average over all data sets [134]. The corresponding threshold is set to 0.54 based on the obtaining results in Figure 6.30 in Chapter 6. It means, the proportional occupancy rate of each cluster to the total number of detected segment of each bird species is compared with the given threshold. Once again the value of parameter m (number of mixture components for additional single model) was set to 20, in order to keep the same complexity recognition system in comparison to the baseline system. The obtained results of this experiment, for the varying test utterance length, between 1 to 3 seconds, are illustrated in Table 7.3.

Table 7.3: Bird species recognition accuracy obtained by baseline system and the proposed element-based system with two different configurations.

Utterance length (sec)	Recognition accuracy (%)		
	Baseline System (Single HMM) n=60	Element-based HMM	
		(n=40, m=20)	(Average number of individual models is 40, m=20)
1	76.02	86.33	85.45
2	82.19	91.01	90.13
3	85.25	93.44	92.42

Table 7.4: Bird species recognition accuracy and error rate reduction obtained by the baseline single and individual element HMM-based recognition system.

Utterance length (sec)	Recognition accuracy (%)		Error Rate Reduction (%)
	Baseline System (Single HMM)	Element-based HMM	
1	76.02	86.33	39.32
2	82.19	91.01	44.58
3	85.25	93.44	48.31

As Table 7.3 shows, there is no further improvement found for this proposed recognition system with the new performed parameters set up. Therefore, the best accuracy rate for utterance lengths of 2 and 3 seconds, resulted from the proposed element-based system; achieved by using 40 individual element models for each bird species and 20 as the number of mixture components per state for single HMM (see Table 7.2). For the 1 second long utterance the best result was obtained for the two parameters set up $\begin{cases} n = 40 \\ m = 20 \end{cases}$ and $\begin{cases} n = 50 \\ m = 10 \end{cases}$, as both recognition values were considered equally (see Table 7.2).

Finally, as Table 7.4 shows, in all cases of using the proposed individual element models there was significant recognition accuracy improvement, with the error rate reduction between 39.32% and 48.61%.

7.4 Review of the recent state of the art in bird species recognition systems

In the following sections, a brief review of previous studies on bird vocalisation recognition systems is presented, a long with presenting a full detailed table of review of these recent state of the art in bird species recognition systems (see Table 7.5).

7.4.1 Data and segmentation procedure

Automatic bird species recognition, based on their vocalisation sounds, has been the subject of many relatively recent studies in the last two decades, dating back to McIlraith and Card [10] in 1997. The data used in many studies up to date, consists of small number of individual bird species [7, 8, 26, 10, 14] with several continuous recordings and nearly all collected from isolated bird vocalisations without noise [7, 8, 10].

In some studies segmenting the continuous recording files into smaller vocalisation segments is performed manually by human intervention of spectrograms [7, 8, 4, 9, 14], or automatically by using an energy-based threshold decision in time or time frequency domain [10, 26, 12, 47, 48, 140, 141]. However, using such an energy-based segmentation method on the data that are recorded in different non-stationary noisy natural environments may obtain low classification accuracy, due to the existence of various background noises or vocalisations of other birds or animals. Instead, some studies [2, 9, 13, 11, 50] used the specific sinusoidal decomposition approach that was proposed in [49], as a manner of automatic segmentation procedure.

McAulay et al. [49] introduced a sinusoidal representation method for speech analysis to obtain a set of sinusoidal components for the speech signal. In the proposed method, first a set of time domain sinusoidal wave segments were detected, then Short Time Fourier Transform (STFT) was applied on the each frame of the corresponding segment. Finally the sinusoidal components were estimated from each spectrum by using the specific peak-picking algorithm. However, the proposed spectral peak picking in [49], manages to find a high number of false sinusoids. In [6, 5] the authors introduced an advanced method for the sinusoidal detection based on the probabilistic modelling of the spectral magnitude pattern and phase continuity around each detected spectral peak. Furthermore, their proposed method is able to use on both stationary and non-stationary sinusoidal signals.

7.4.2 Signal representation and modelling

For the purposes of feature extraction, a large number of previous studies were influenced by common STFS-based feature representations that are employed in the field of speech signal processing, such MFCCs which are used in bird recognition studies [7, 26, 50, 2, 51, 52, 141, 142, 143, 144] and LPCCs [7, 50, 9]. As the general form of MFCCs is capable of storing the entire frequency range, they are likely to capture the environmental background noises along with the other frequency bands corresponding to other concurrent bird vocalisation in the data. Some studies such as [2, 9, 13, 4, 145] employed a series of spectral statistic frames that were capable of characterising each obtained spectro-temporal segment. Since this feature representation is in the form of a single dimension vector, these features may not be able to present the syllables that have more complexity in pattern and also could be sensitive to any variations in segmentation. A few recent studies [6, 48, 2], including this thesis, employed the segments which are obtained in the segmentation step by performing sinusoidal detection as a means of a temporal sequence of frequencies features (which are referred to as frequency track in this thesis). These frequency track features are powerful to deal with the various background noises and often also other concurrent birds/animals' vocalisation, which exist in the real naturally recorded field data. The performance of using the frequency track, which was obtained based on the proposed method in [6, 5], for the recognition of tonal bird sounds in a noisy environment is demonstrated in [3], with further comparison to MFCC features. The obtained result in [3] illustrates that the accuracy of the recognition system based on the frequency track features has been significantly improved, compared to the MFCC-based system in noisy background conditions.

The most commonly used modelling approaches include dynamic time warping (DTW) [8, 7, 9], Gaussian mixture modeling (GMM) [2, 26, 140, 146, 14, 143], hidden Markov models (HMMs) [7, 2, 48, 9], support vector machines (SVM) [12, 142, 141], and Artificial neural networks (ANNs) [10, 147, 47, 51].

All the relevant information about the data set, segmentation procedure, signal representations (features), classification methods and the best achieved recognition results of previous best studies in bird sound recognition systems is reviewed in more detail, along with this proposed novel element-based recognition approach, in Table 7.5. The main application of the studies which are referenced in Table 7.5, is to develop a bird species identification system, unless otherwise stated in the note's section.

Table 7.5: Previous works in bird sound recognition systems

Papers	Data set used	Segmentation	Feature(s)	Classification method(s)	Best achieved result
This thesis	48 bird species (38 hours in total)	Automatic	Frequency tracks	HMM	93.44% (Element-based HMM)
[8]	2 bird species	Manual	Spectral signal	DTW	97% (for song-based system)
Note: The proposed method directly compared the spectrograms of input bird sounds with a set of manually predefined templates, corresponding to each class of recognition. The authors applied this method to small vocalisation data from two bird species, recorded in a low-noise environment; 97% and 84% accuracy was achieved, respectively, for the song-based and syllable-based methods. The proposed method did not use amplitude normalization, so the results may be sensitive to amplitude differences.					
[10]	6 bird species (133 songs in total)	Automatic	Short-time spectrum of the signal	ANNs	93%
Note: This study was among the first to apply automatic classification to a larger number of bird species (in total 6 bird species). In the proposed method the bird signal was represented with spectral and temporal parameters of the songs. The classification accuracy is 82% using a backpropagation neural network for identification and 93% using quadratic discriminant analysis.					
[7]	2 bird species	Manual	MFCCs vs LPCs	MFCC-HMM vs LPC-DTW	Above 95% (for both proposed methods)

Papers	Data set used	Segmentation	Feature(s)	Classification method(s)	Best achieved result
Note: In this study, the performance of two classification methods, HMM and DTW, is compared for automated recognition of bird song units from continuous recordings. Each recording was segmented into a set of syllables' signals, manually. The method was based on template matching of spectrograms of these segments.					
[11]	4 Bird species	Automatic	Spectral features (based on sinusoidal modelling of syllables)	-	-
Note: The authors proposed an approach, related to the development of techniques for automatic recognition of bird species, to classify the bird sounds into four classes based on their harmonic structure. Each harmonic component was modelled with one time-varying sinusoid. The features' extraction step was performed by using a parametric line spectrum estimation method, to obtain a sinusoidal model for the syllables. Actual recognition results have not been reported in the current article, but the study shows that the use of a time-varying sinusoidal model instead of the centre frequency of a syllable, improves the accuracy of the song recognition rate by 10-30 %.					
[147]	289 bird species (2400 audio files in total)	Automatic	A subset of spectral features	Artificial neural networks (ANNs)	73% (obtained with seven features)
Note: In this study, the syllable segmentation was done by means of short-time signal energy and the short-time maximum of the spectrum. Also, each syllable is represented with 19 low-level acoustical features: seven features measuring the short-time behaviour of the syllable; followed by the mean and variance of the feature trajectories; and lastly, five features to describe static properties over the entire syllable. The best overall recognition result was 73%, obtained with seven features. However, it seems, that the number of parameters is probably too high for an efficient recognition algorithm.					

<table>
<tr><th>Papers</th><th>Data set used</th><th>Segmentation</th><th>Feature(s)</th><th>Classification method(s)</th><th>Best achieved result</th></tr>
<tr><td>[2]</td><td>14 bird species</td><td>Automatic</td><td>MFCCs vs a vector of various descriptive features</td><td>GMM vs HMM</td><td>71.3% (for MFCC-GMM system)</td></tr>
<tr><td colspan="6">Note:
In this study, the segmentation of a recording into individual syllables is performed using an iterative time-domain algorithm [11]. The recognitions were performed based on two methods: the use of single syllables and song fragments. The single-syllable based recognition accuracy was around 40% and using song-based recognition increased the accuracy rate to around 70%. There were only small differences between the results of the GMMs and HMMs in general. Also, the authors described that the results with song level parameterisation are significantly better than the results from the recognition of single syllables.</td></tr>
<tr><td>[9]</td><td>12 bird species (20 sound files per bird species)</td><td>Manual</td><td>MFCCs, LPCCs and a set of spectral features by time-variant analysis (spectral peak tracks)</td><td>DTW, HMM and SPT</td><td>99% (for SPT)</td></tr>
<tr><td colspan="6">Note:
This article compares the performance of three recognition systems on the same bird database: LPCC-DTW based (90% correct matches for clean data and 71% for noisy data), MFCC-HMM based (95% correct matches for clean data and 76% for noisy data) and Spectral peak track (SPT) method (99% correct matches for the entire natural and synthetic database at a high SNR). In the segmentation stage, the segment is assumed to be edited manually to contain a suitable segment for the matching process.</td></tr>
<tr><td>[50]</td><td>Two noise free datasets: 420 and 561 bird species (one recording file for each species)</td><td>Automatic</td><td>LPCCs vs MFCCs</td><td>Linear discriminant analysis (LDA)</td><td>87% (MFCC-LDA system)</td></tr>
</table>

<table>
<tr><th>Papers</th><th>Data set used</th><th>Segmentation</th><th>Feature(s)</th><th>Classification method(s)</th><th>Best achieved result</th></tr>
<tr><td colspan="6">Note:
In this study, each input signal is first segmented into a set of syllables using the automatic method proposed by [11]. Also, for the purpose of modelling, the authors used a codebook consisting of several representative feature vectors to model the detected syllable segments.</td></tr>
<tr><td>[26]</td><td>11 bird species</td><td>Automatic</td><td>MFFCs</td><td>GMM</td><td>90%</td></tr>
<tr><td>[145]</td><td>15 bird species (each bird has recordings ranging from 3 to 7 in number)</td><td>Automatic</td><td>MFCCs vs the average spectrum over time</td><td>DTW, GMM and SEAV</td><td>87% (SEAV)</td></tr>
<tr><td colspan="6">Note:
This study introduced a new representation for bird vocalisation syllables, which is based on the average spectrum over time, for identification of bird calls. Each representation feature is calculated on the FFT spectrum of each bird and is called the spectral ensemble average voice print (SEAV) of that particular bird. The SEAV classification method was based on template matching and it achieved the best accuracy over other reference recognition systems (DTW and GMM based approaches).</td></tr>
<tr><td>[47]</td><td>8 bird species</td><td>Automatic</td><td>Wavelet coefficients</td><td>ANNs</td><td>96% (for MLP system)</td></tr>
<tr><td colspan="6">Note:
In this study, each detected segment (syllable-based) was presented with four parameters derived from a wavelet decomposed signal representation. Then, these features were used as inputs of two neural networks: the unsupervised self-organizing map (SOM) and the supervised multilayer perceptron (MLP). The results showed that the SOM network recognized 78% and the MLP network 96% of the test sounds correctly.</td></tr>
</table>

<table>
<tr><th>Papers</th><th>Data set used</th><th>Segmentation</th><th>Feature(s)</th><th>Classification method(s)</th><th>Best achieved result</th></tr>
<tr><td>[12]</td><td>8 bird species</td><td>Automatic</td><td>MFCCs vs a set of low-level signal parameters (11 low level descriptive parameters)</td><td>Support vector machine (SVM)</td><td>97% (for MFCC-based system)</td></tr>
<tr><td colspan="6">Note:
The segmentation of a recording into individual syllables is performed using an iterative time-domain energy based detector method, which is presented in [147]. It seems, however, that the number of parameters is probably too high for an efficient recognition algorithm.</td></tr>
<tr><td>[48]</td><td>9 bird species (520 calls segments in total)</td><td>Automatic</td><td>The peak frequency and short-time frequency bandwidth</td><td>HMM</td><td>84% (overall for bird)</td></tr>
<tr><td colspan="6">Note:
This study presents an automatic call recognition system for birds, crickets and frogs that have a narrow short-time frequency bandwidth in their vocalisation. Also, it proposed an approach to extract vocalisation signals from background noise using a frequency band threshold filter on spectrograms. It is concluded that the performance of the above process is sensitive to the threshold-band filtering step.</td></tr>
<tr><td>[13]</td><td>4 bird species (includes about 200 short vocalisation signals)</td><td>Automatic</td><td>Frequency track sets</td><td>Mahalanobis distance function</td><td>79%</td></tr>
<tr><td>[51]</td><td>420 bird species (one recording file for each species)</td><td>Automatic</td><td>MFFCs</td><td>ANNs</td><td>65% (overall)</td></tr>
</table>

Papers	Data set used	Segmentation	Feature(s)	Classification method(s)	Best achieved result
Note: The amount of data for each bird species is very low (only one recording per bird species).					
[14]	4 bird species	Manual	MFCCs	GMM	90.45% (overall)
[140]	94 bird species (four bird song signals per bird species)	Automatic	MFCCs	GMM, Note n-gram modelling	89.5% (overall for system combination)
Note: This study introduced a bird species' identification system, by using GMM and a universal background model (GMM-UBM) on a closed set. The segmentation was performed using a simple VAD system to extract bird vocalisation segments from background signals.					
[4]	13 bird species (90 minutes of data in total)	Manual	A set of spectral features (e.g. min-frequency, max-frequency and bandwidth)	MIML	96.1%
Note: This study presents a bird sound detection system by using a multi-instance multi-label (MIML) framework. The segmentation procedure was done manually.					
[142]	75 bird species	Automatic	MFCCs	SVM	59.75% (overall)
Note: This study presents a method that deals with the segmentation step of the audio signal in the automatic bird species' identification problem. The authors mainly focused on comparing the recognition results of their proposed automatic segmentation system with a manual segmentation system. The obtained results showed that using an automatic segmentation method improves the overall recognition accuracy from 52.78% to 59.75%.					

<table>
<tr><th>Papers</th><th>Data set used</th><th>Segmentation</th><th>Feature(s)</th><th>Classification method(s)</th><th>Best achieved result</th></tr>
<tr><td>[144]</td><td>10 bird species</td><td>Automatic</td><td>Permutation Coefficients (PCs) and MFCCs</td><td>k-NN</td><td>65% (using PCs)</td></tr>
<tr><td colspan="6">Note:
The main focus of this article is introducing a new parametric representation (refers to permutation coefficients) of bird sounds for automatic identification of their species. The method is based on the distribution of short temporal patterns in bird vocalization. The classification was done using a k-Nearest Neighbours (k-NN) method.</td></tr>
<tr><td>[143]</td><td>40 bird species (5.3 hours in total)</td><td>Automatic</td><td>MFCCs</td><td>GMM</td><td>71.5%</td></tr>
<tr><td colspan="6">Note:
This study proposed a robust frame selection for bird species' recognition. Only best frames that represent the dominant sounds were selected and parametrized by MFCCs.</td></tr>
<tr><td>[141]</td><td>30 bird species</td><td>Automatic</td><td>MFCCs vs MWSC</td><td>SVM</td><td>85.25% (for MWSCC-SVM)</td></tr>
<tr><td colspan="6">Note:
This paper proposed a new bird sound classification approach based on adaptive energy detection, to improve the recognition accuracy of bird vocalisation in noisy environments. The proposed approach extracted two types of feature representations: MFCC and mel-scaled wavelet packet decomposition sub-band cepstralc Coefficient (MWSCC). The results show that the MWSCC-based representation has a better noise immunity function, and the recognition performance was improved significantly, by using the proposed adaptive energy detection.</td></tr>
<tr><td>[146]</td><td>127 recording files from 15 individual male-chiffchaff bird species (about 11 hours in total)</td><td>Automatic</td><td>Linear Frequency Cepstral Coefficients (LFCCs)</td><td>GMM-UBM</td><td>78.5% (overall)</td></tr>
</table>

Papers	Data set used	Segmentation	Feature(s)	Classification method(s)	Best achieved result
Note: This article presents an automatic system for recognition of bird individuals based on a GMM and a universal background model (GMM-UBM) method, extended by an advanced voice activity detection (VAD) algorithm. The overall achieved identification accuracy is 78.5% . The above proposed approach was only tested with 15 individual males of one bird species (chiffchaff), whereas in this study, the performances over multiple species are of interest.					

As can be seen in Table 7.5, each system was performed with different data sets, segmentation, features and modelling/classification approaches. The studies which employed manual segmentation [7, 8, 4, 9, 14], obtained a very good recognition accuracy of more than 90%. However, employing manual segmentation is not of interest in this section, as the aim of this review is to compare the result of this proposed element-based approach with different developed automatic systems for the identification of bird species. Hence, the best achieved accuracy rate, over all automated approaches referenced in Table 7.5, ranges between 59.75%-97% (on average 81%). Compared to these recognition rates, the experimental results of the proposed novel element-based recognition system demonstrate a very high accuracy rate of 93.44%. There are only a few automated identification systems stated in Table 7.5 which obtained a higher recognition rate (more than 94%) [9, 12, 47]. However, these systems are tested with a small amount of data or number of bird species. In general, most of the mentioned studies used a data set with less than 15 bird species, and only some studies [147, 50, 51, 140, 142, 143, 141] used data from more than 30 bird species. While this proposed approach is evaluated with a data set that consists of 48 bird species, with an overall time of 38 hours and recorded in real-world natural environments.

7.5 Summary

In this chapter, two automatic bird sound recognition systems were developed. First, a baseline system is presented that used only a single HMM for each bird species. Secondly, a novel element-based system is presented that employed a set of HMMs to model the individual elements. In the second system, the clustering outcome of the unsupervised discovery of the bird vocalisation approach (as presented in Chapter 6), was used as the input element-level labelling information. All the proposed recognition systems are based on the frequency track segments which were detected by employing the described sinusoidal detection and segmentation approach in Chapter 5. The temporal frequency track and the corresponding dynamic derivative features (delta and acceleration) were modelled by using hidden Markov models (HMMs). Experimental evaluations were performed on a total of 48 bird species' subsets; the entire data lasts about 38 hours, from several field recordings. The experimental results demonstrated that the proposed individual element HMM-based system provided over 39% bird species recognition error rate reduction, compared to the single HMM-based system of the same complexity.

Chapter 8

SUMMARY AND FUTURE WORKS

8.1 Introduction

This chapter summarises and concludes the outcome of this research and also includes novel contributions. Furthermore, recommendations for further studies in this field of research are included in this chapter.

8.2 Major Contributions

The research introduced in this thesis provides original contributions to the field of automatic processing and classification of bird acoustic signals. The major contributions of this thesis are summarised as follows:

1- The manual annotation of a large corpus of bird vocalisation recording files, which are recorded in natural field environments. In order to evaluate the performance of the automatic segmentation and feature extraction procedures, and also to allow other researchers to perform and evaluate comparative experiments with the real field corpus, the whole data is inspected manually to obtain the acoustic event-based label file for each vocalisation recording. For the sake of refining or modifying the acoustic events in the provided labelling files, a user interface (a MATLAB script) has been supplied among the annotation files (Chapter 4).

2- Development of a novel discovery approach in an unsupervised way that can find a set of individual vocalisation elements for each bird species, based on the detected segments in the data. This approach is performed in two consecutive stages:

- Development of a partial DTW similarity calculation algorithm to search for the partial and multiple matchings between a given pair of (temporal) sequences. In order to find the partial similarity paths, the proposed algorithm employs a variant of DTW in several searching procedures; where each search considers a different time-stamp on one of the sequences and allows the DTW alignment path to start and end anywhere on the second sequence. The obtained pairwise partial similarity path can be represented by the partial similarity score and the corresponding time-stamps of the detected partial path on both sequences. This stage is explained in Chapter 6 (section 6.2).
- Development of a novel hierarchical clustering algorithm that employs the (obtained) partial similarity information, including the similarity scores and the corresponding time-stamps of the partial similarity path, of the entire vocalisation segments of each bird species in order to group all the homogeneous structured (detected) segments into a set of distinct element-based vocalisation clusters. Several rules and conditions are used in this method to control the merging decisions. In other words, the merge

decisions of the clustering procedure are always based on further investigations of the prospective likeness structure of the group, including both the merging objects. The resulting outcome of this novel approach offers the label information for each discovered segment of the data, and is then used in Chapter 7, as the element-level vocalisation labelling information, to train the (HMM) models. This stage is explained in Chapter 6 (section 6.3.2).

3- Development of a novel automatic bird species' identification system based on HMM modelling of individual element vocalisation units. In this approach, instead of employing a single HMM model for each bird species, a single HMM is used to model each type of vocalisation pattern that is available in each particular bird species. As there is no further element-level information available among the natural field recordings, training the element-based models is not practical. Hence, this proposed system employs the outcome of a hierarchical clustering algorithm, as label information, to train the HMM models. This novel approach is explained in Chapter 7.

8.3 Summary

This thesis aimed to present an automatic system for the identification of bird species from natural field recordings. In the first part of this thesis (Chapter 3), a literature review of the biological theories of bird vocalisations by studying the communication and singing behaviours of typical passerine birds is presented. This is followed by a description of the corresponding bird vocalisation terminology, as in songs, calls, syllables, phrases and elements; and finally, the scientific theories about vocal learning and development procedures in young birds are discussed.

Then in the next chapter (Chapter 4), a summary of the large, available bird vocalisation archives is presented, followed by a description of a smaller data set which has been used in recent bird classification challenges. Based on this literature, a large data set from the natural field consisting of 50 tonal bird species with over 900 recording files were collected from the Borror Laboratory of Bioacoustics' archive of bird vocalisation. Each corresponding recording file of this vocalisation corpus was originally labelled with the name of the particular bird that produced the main vocalisations in the file. This original labelling was used in all the provided experimental evaluations in this study. Meanwhile, as the data had been recorded in the real habitats of birds, each recording also contained some irrelevant audio information along with the desired bird vocalisation signal. In order to allow other researchers to perform and evaluate comparative experiments with the real field corpus, each recording of the data was annotated manually into set of pre-defined audio sub classes.

In order to develop an automatic bird sound recognition system, for the purpose of automatic segmentation and feature extraction tasks, by using the proposed sinusoidal detection approach in [5, 6], each recording file of the data was extracted and separated into the set of distinct peak frequency components (frequency track segments) to characterise bird tonal vocalisation (Chapter 5). These obtained segments are used as input sequences for the further processing of this research.

HMM was employed in this research for training the models and classification purposes. Hence, two automatic HMM based recognition systems, baseline and element-based, were developed in this thesis for the identification of bird species from natural field recordings (Chapter 7). Both systems are bird species' recognition based. For the base line system the entire training frequency tracks' segments of each bird species were modelled with a single HMM model with 13 HMM states. Each state was modelled by a mixture of Gaussians, to allow for the collection of element patterns and the diversity of individual entities of vocalisations. For the second proposed system, it was the aim to build a separate HMM to model each type of bird vocalisation pattern for each bird species, instead of employing a single model for each bird species. As there is no further element-level information available among the natural field recordings, training the element-based models is not practical. To deal with this issue, an approach for unsupervised discovery of acoustic elements in bird vocalisations is presented in this research (Chapter 6).

This approach is accomplished in two consecutive stages. In the first stage (section 6.2), unlike a conventional dynamic time warping (DTW) algorithm which calculates the similarity of whole sequences, the presented modified DTW algorithm allows searching for multiple matches, possibly partial, between each pair of detected segments (frequency tracks segments). The outcome of the DTW search is a set of found partially matching paths, with their corresponding similarity values. These similarity values are called partial similarity scores, as they were obtained during the searching for partial paths. Each partial similarity score is calculated based on a combination of the cumulative distance of the DTW path match, the length of the matching path and the ratio of the length of the matching path to the total length of the segment. In the next stage (see section 6.3.2), the outcome of the DTW searches, including the similarity distance matrix and the associated time-stamps information of the obtained partial similarity paths, were then used in a novel proposed hierarchical clustering approach, to group all the homogeneous structured segments into a set of distinct element-based vocalisation clusters. Several joining rules and conditions were employed in the proposed clustering method to control the joining procedures. The merge decisions were always based on further investigations of the prospective likeness structure of the group, including both the merging objects. The experimental evaluations in Chapter 6 demonstrated that the obtained clusters showed good coherence and provided a set of structurally distinct bird vocalisation patterns.

Finally, the resulting outcome of above discovery approach offers the label information for each discovered segment of the data. Hence these label information were used to train the HMMs in the novel proposed element-based recognition system. Experimental evaluations were performed on a total of 48 bird species' subsets; the entire data lasts about 38 hours, from several field recordings. The experimental results demonstrated that the proposed individual element HMM-based system obtained a recognition accuracy of over 93% by using 3 seconds of detected signal and over 39% recognition error rate reduction, compared to the baseline HMM system of the same complexity.

8.4 Future research directions

Perpetually there will be a countless number of ways in which the current work can be extended. In the following, some of the possible extensions and future works are expressed.

- The current corpus data are collected from Borror's bird vocalisation archive with its own quality categories between 'fair to good' and 'very good' in Table 4.1. These collected data are relatively clean, in comparison with the real world which contains many more background noises. Hence, it is suggested to extend the corpus data with vocalisation audio files that are recorded in a noisier natural environment.
- More experts are needed to annotate the vocalisation recording files.
- In the partial DTW similarity calculation method the following suggestions are made to overcome the current limitations:
    - Modification of the DTW function by incorporating the other constraints on the warping path, e.g. constraints on the slope variation of the cumulative path, to improve the partial similarity detection.

  - In the case of obtaining multiple partial matching paths after the boxing procedure (section x), the corresponding segments may consist of several repetitions. By further investigation among the segments, i.e. the gaps between paths, the corresponding segment can split into a few smaller segments, if possible.
  - Automatic estimation for the value of the minimum length of a cumulative path (L_{min}) for each bird species separately, instead of using a fixed value for all the bird species.

- In the proposed agglomerative hierarchical clustering method, the following suggestions are available to improve the clustering outcome by detecting the the least number of mismatched segments in each vocalisation cluster:
  - Automatic estimation for the value of parameters thr_{drop} and D_{thr} (in section 6.3.4) for each bird species separately, instead of using a fixed value for all the bird species.
  - Optimizing the whole clustering procedure in order to obtain a smaller number of clusters with a higher occupancy of elements, for each bird species.

- Modifying the proposed recognition system in order to deal with the multiple species in the given utterances of the test signals.

- Modifying the proposed recognition system in order to deal with the test utterances that contain the vocalisation information of a particular bird species which is not listed in the system.

- Employing the outcome of the manual annotation procedure to build separate training models for each obtained audio sub-class, to deal with irrelevant background sounds and noises.

- Developing the proposed identification system by using the discriminative classification approaches, such as: Support Vector Machines and Artificial Neural Networks.

- Speeding up the computational demands for the model training procedure and the partial DTW searching procedure.

Appendix A

LIST OF AVAILABLE BIRD SPECIES IN CORPUS DATA

Bird #	Bird Name	Number of File	Data Length (Minutes)	Num Groups
Bird 1	CarolinaWren	19	95	285
Bird 2	IndigoBunting	33	89	366
Bird 3	LarkSparrow	10	56	255
Bird 4	CanadaWarbler	15	28	177
Bird 5	ChippingSparrow	17	31	245
Bird 6	FoxSparrow	14	52	146
Bird 7	HermitThrush	20	55	257
Bird 8	HouseFinch	24	45	355
Bird 9	LouisianaWaterthrush	15	36	96
Bird 10	NashvilleWarbler	16	34	124
Bird 11	NorthernWaterthrush	13	33	151
Bird 12	PineWarbler	38	54	176
Bird 13	PurpleFinch	18	42	348
Bird 14	BaltimoreOriole	22	51	70
Bird 15	CommonYellowthroat	13	38	88
Bird 16	EasternMeadowlark	17	34	85

Bird #	Bird Name	Number of File	Data Length (Minutes)	Num Groups
Bird 17	EasternWoodPewee	31	55	85
Bird 18	GrayCatbird	18	48	288
Bird 19	GreenTailedTowhee	19	65	238
Bird 20	HoodedWarbler	11	41	138
Bird 21	HouseWren	11	48	533
Bird 22	MarshLlongBilledWren	17	50	270
Bird 23	NorthernCardinal	25	93	103
Bird 24	Ovenbird	18	43	136
Bird 25	RoseBreastedGrosbeak	24	56	231
Bird 26	ScarletTanager	26	49	151
Bird 27	SummerTanager	21	46	147
Bird 28	SwampSparrow	26	45	106
Bird 29	VesperSparrow	44	63	303
Bird 30	YellowWarbler	21	46	186
Bird 31	ProthonotaryWarbler	12	32	88
Bird 32	OliveSidedFlycatcher	13	37	28
Bird 33	MagnoliaWarbler	32	47	191
Bird 34	KirtlandsWarbler	18	40	79
Bird 35	KentuckyWarbler	19	34	66
Bird 36	AmericanGoldfinch	34	47	272
Bird 37	AmericanRedstart	25	42	182
Bird 38	CarolinaChickadee	9	33	24
Bird 39	BlueGrosbeak	13	34	158
Bird 40	WilsonsWarbler	21	41	169
Bird 41	WhiteEyedVireo	10	61	133
Bird 42	WarblingVireo	15	30	216
Bird 43	SavannahSparrow	12	34	131
Bird 44	NorthernYellowShaftedFlicker	27	32	109
Bird 45	FieldSparrow	19	32	84
Bird 46	SlateColoredJuncoDarkEyed	27	32	175
Bird 47	WillowFlycatcher	11	29	103
Bird 48	WinterNorthernWren	9	42	209
Bird 49	WesternMeadowlark	8	57	87
Bird 50	YellowThroatedWarbler	14	33	123

Appendix B

LIST OF AVAILABLE BIRD SOUND FILES IN CORPUS DATA

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
Bird 1: CarolinaWren			
2418	5	0	03:20
6133	5	0	03:40
6175	5	0	04:10
6206	5	0	08:07
8876	5	0	02:19
10965	5	1	03:11
12240	5	0	04:08
13309	5	0	05:31
13402	5	0	15:42
13893	5	0	05:51
14005	5	0	15:49
14164	5	1	10:51
15713	5	0	02:58
15722	5	0	03:10
15871	5	0	01:19
15897	5	0	02:00
16375	5	0	01:25
16377	5	0	01:12

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
24088	5	0	00:31
Bird 2: IndigoBunting			
1466	5	0	00:58
1982	5	1	01:49
2607	5	1	01:43
3349	5	1	01:50
3392	5	0	02:43
5152	5	0	03:20
5201	4	1	06:43
6957	5	0	01:42
7478	5	0	01:24
8318	3	3	04:45
8319	3	3	02:15
8320	4	2	03:03
8353	5	1	01:36
8373	4	3	04:55
8389	5	1	02:45
8392	5	1	03:36
8393	5	2	04:19
8395	5	1	04:09
8396	5	0	02:12
8558	5	0	02:04
8642	5	1	03:52
9489	5	0	01:19
13556	5	0	01:25
14329	5	1	06:37
14796	5	0	01:42
14824	5	0	01:16
15634	5	0	00:52
16387	5	0	01:54
17086	5	1	01:28
22588	5	0	00:22
30114	5	1	05:29
30119	5	0	02:23
30124	5	0	02:18
Bird 3: LarkSparrow			
3384	5	0	16:11
3385	5	2	01:44
3386	5	2	09:22
4402	5	0	04:40
4418	5	1	03:31
5536	5	1	02:02
5580	5	0	02:38

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
5596	5	1	05:12
8563	5	2	06:51
30814	4	1	04:00
Bird 4: CanadaWarbler			
1060	5	0	01:18
1460	5	1	01:50
2668	5	0	01:33
2909	5	1	01:48
3443	5	0	03:57
3445	5	1	01:14
4037	5	0	02:12
4082	5	1	03:24
5213	5	0	01:41
5311	5	0	02:01
6845	5	0	01:45
6907	5	0	00:44
7533	5	0	01:56
11838	5	2	01:25
14789	5	0	01:42
Bird 5: ChippingSparrow			
7542	5	0	01:02
9941	5	0	03:23
10025	5	0	01:46
10895	5	1	01:18
11113	5	0	01:05
12376	5	0	01:54
14102	5	0	01:12
15971	5	2	01:27
16050	5	0	02:15
16423	5	1	01:09
16464	5	0	01:15
16468	5	2	01:14
17055	5	0	02:40
18728	3	4	05:33
22191	5	0	01:28
22192	5	2	01:16
24426	4	2	01:13
Bird 6: FoxSparrow			
1664	3	1	00:13
18806	5	0	03:57
18829	4	3	04:55
23974	5	1	02:34
23981	5	0	02:23

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
24022	5	1	06:14
24560	4	0	04:20
24941	5	2	03:51
26408	5	1	15:00
27843	5	2	01:54
28482	4	2	02:17
28459	5	0	02:12
29975	5	0	01:30
34995	4	3	00:44
Bird 7: HermitThrush			
619	5	1	01:34
1551	5	0	02:08
2184	5	0	04:15
2206	5	0	03:48
2214	5	0	03:45
2217	5	0	03:49
2233	5	1	01:53
2942	4	1	01:48
3001	5	1	02:09
3553	4	1	02:35
4375	3	1	04:49
7862	5	1	01:43
9199	5	0	02:12
16950	5	0	04:20
20571	4	2	02:06
22657	5	0	00:24
24025	5	1	05:54
29046	5	0	01:22
29047	5	0	03:50
32658	5	1	01:16
Bird 8: HouseFinch			
7080	5	0	00:48
7129	5	0	02:34
10115	5	1	01:07
10186	5	0	00:48
10256	5	0	00:51
11581	5	0	02:04
11591	5	0	00:58
11613	5	0	01:50
11892	5	1	02:37
12380	5	0	02:02
16447	5	0	00:12
16451	5	0	02:19

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
16804	5	0	05:25
16866	5	1	01:49
16903	5	0	00:30
17548	5	0	01:19
18892	5	1	04:00
24085	4	3	03:47
24977	5	0	01:04
25293	5	1	02:15
26098	4	1	00:35
26404	5	0	04:09
28901	5	1	00:54
33630	3	2	01:09
Bird 9: LouisianaWaterthrush			
796	5	0	03:44
1691	5	1	04:45
1924	5	0	00:39
4537	5	0	02:00
5099	5	0	02:06
5814	5	0	02:26
7421	5	0	01:28
12129	5	0	00:38
12134	5	0	02:00
12656	4	2	04:06
12682	4	3	03:51
12823	5	0	01:02
15773	5	0	01:48
21888	5	0	01:45
24421	5	0	03:33
Bird 10: NashvilleWarbler			
1396	5	0	02:20
2113	5	0	01:12
3336	5	0	01:48
3487	5	0	02:13
4209	5	0	02:05
4588	5	0	01:10
4756	5	0	01:42
5761	5	1	02:09
15620	5	0	02:00
15883	4	1	02:15
18692	4	1	03:29
18827	5	0	00:57
28298	5	0	01:29
28321	5	0	02:42

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
29220	5	0	04:34
34924	5	1	01:58
Bird 11: NorthernWaterthrush			
1418	5	0	02:36
3936	5	0	01:35
4636	5	0	03:21
8260	5	0	02:55
9473	5	1	01:20
10569	5	1	01:35
12592	4	2	03:18
12593	5	1	02:44
12596	5	0	02:36
12618	5	1	04:48
12619	5	1	03:13
17554	5	1	01:12
23963	5	0	03:00
Bird 12: PineWarbler			
2115	5	0	02:00
3775	5	1	02:48
4202	5	0	02:18
4429	5	0	01:24
4450	5	2	01:58
4469	5	0	01:04
4966	4	3	01:09
5028	4	2	02:07
5717	5	0	01:50
5730	5	0	01:42
8371	5	0	02:31
9042	5	0	01:24
9347	5	1	00:40
10550	5	0	02:12
11100	5	0	02:13
11935	5	1	00:49
11939	5	0	00:57
13156	5	0	00:31
13621	5	0	00:24
13853	5	0	00:06
14083	5	0	00:20
14092	5	0	01:02
14741	5	0	00:45
14748	5	0	00:57
14753	5	0	01:04
15059	5	1	01:20

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
15469	5	1	02:08
15602	5	1	00:42
15716	5	0	01:43
15725	5	0	01:18
16232	4	2	00:59
16239	5	0	00:54
16452	5	0	00:09
16555	5	0	00:18
16696	5	1	01:10
16706	4	3	02:47
26423	2	4	03:19
26424	3	3	03:28
Bird 13: PurpleFinch			
1556	5	0	01:07
1876	3	1	02:40
2875	5	0	02:18
3371	5	2	01:20
3467	5	0	03:32
3621	5	1	01:12
3676	5	0	02:08
3874	5	1	02:15
3890	5	1	01:31
4858	5	1	02:18
6948	5	0	01:47
6218	4	2	02:15
7735	5	0	02:16
7874	5	0	01:39
8303	5	1	02:03
9284	5	0	03:10
11112	5	1	03:28
29677	5	1	05:29
Bird 14: BaltimoreOriole			
503	5	1	01:13
515	4	2	01:02
527	4	3	02:13
1446	5	1	01:34
2798	5	0	03:05
3340	4	3	02:22
5129	5	0	01:19
5766	5	0	01:04
5822	5	1	00:15
7499	5	1	01:27
7540	5	2	04:41

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
7579	5	3	02:36
8935	5	1	02:25
9471	5	2	02:48
11209	5	1	02:20
11383	5	1	04:14
12184	5	1	02:35
12220	5	1	02:47
12894	5	1	03:05
15828	5	1	02:12
16311	5	1	02:10
30234	5	1	03:18
Bird 15: CommonYellowthroat			
7556	5	0	04:54
7885	5	0	01:27
9052	5	0	03:18
9387	5	0	01:29
9738	5	0	02:30
13504	5	0	03:23
14846	5	1	01:31
15798	5	1	02:08
15942	5	2	09:32
16265	5	1	01:57
16400	5	2	01:43
16690	5	1	00:50
17092	5	3	04:04
Bird 16: EasternMeadowlark			
5181	5	0	03:57
5529	5	0	01:07
5751	5	0	02:48
6228	4	2	01:08
6229	5	0	02:15
6253	5	0	01:35
7375	5	1	01:32
9850	5	1	02:09
9984	4	2	03:58
10415	5	0	01:28
11003	5	0	03:15
11018	5	0	02:16
11287	5	0	02:21
12072	4	4	01:23
13386	5	0	01:27
14236	5	1	01:03
17635	4	2	00:41

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
Bird 17: EasternWoodPewee			
1459	5	0	04:26
3452	5	0	02:12
4064	5	1	02:41
4072	5	1	01:53
4084	5	2	01:35
4100	5	1	03:03
5113	5	0	01:15
5175	5	0	01:45
5715	5	0	01:41
6411	5	0	01:52
6857	5	2	01:25
6953	5	1	01:33
7532	5	1	01:41
9004	5	1	01:35
9007	5	0	00:55
10524	5	0	00:48
10609	5	1	01:33
11266	5	0	00:47
12248	4	3	00:30
12262	5	2	00:44
12983	5	1	01:26
13539	4	1	01:10
13704	5	0	05:46
13912	5	2	01:03
14814	3	2	01:48
15710	3	2	01:15
16345	4	2	00:42
17587	3	3	00:56
21529	3	3	01:47
25230	4	1	01:06
29998	5	1	04:22
Bird 18: GrayCatbird			
1371	5	0	02:04
4574	4	0	02:05
6259	5	0	03:06
8187	5	0	02:39
8225	5	0	03:26
10037	5	0	01:18
11191	5	0	01:24
11301	5	1	02:17
12513	5	0	00:44
13128	5	0	02:48

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
13824	5	0	03:48
14856	5	0	03:46
15638	5	0	04:32
15822	5	0	03:05
15881	5	0	04:23
28431	5	1	02:29
28490	5	1	02:45
28495	5	1	02:04
Bird 19: GreenTailedTowhee			
10338	5	0	03:45
10761	5	0	04:18
10764	5	0	03:33
10773	5	0	02:14
11534	5	0	04:23
22787	5	1	01:34
26906	5	1	06:15
26908	5	0	02:58
26909	5	0	01:31
27956	5	0	01:59
28933	5	0	00:54
29100	5	0	02:51
29101	5	0	01:06
29102	5	1	02:56
29569	5	0	00:58
30963	4	3	09:57
30964	3	5	12:45
32310	5	1	00:31
32313	4	3	01:17
Bird 20: HoodedWarbler			
16783	5	2	04:47
16793	5	2	03:03
17064	5	1	02:28
17516	4	3	05:38
17579	5	1	03:04
17593	5	2	01:41
17909	4	2	01:20
17955	4	3	03:20
17947	4	1	09:23
17949	5	1	02:32
17963	5	2	04:14
Bird 21: HouseWren			
12110	5	0	01:59
12844	5	0	01:08

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
16263	4	2	03:18
28579	1	4	00:34
28839	4	3	05:05
29141	5	1	01:03
29511	4	3	10:52
29580	5	1	06:10
30824	4	3	04:22
30825	4	3	06:34
30826	4	3	07:45
Bird 22: MarshLlongBilledWren			
493	4	2	07:00
3413	5	0	03:03
4983	3	3	03:44
4985	3	4	03:13
5979	5	0	01:34
6422	5	1	01:25
7679	5	0	01:15
7774	5	1	01:39
7817	5	0	03:27
9181	5	1	01:55
11361	5	0	02:19
14060	5	0	02:21
14108	5	0	01:27
14689	5	0	01:13
18824	5	0	01:11
28721	5	2	09:54
33979	5	0	03:17
Bird 23: NorthernCardinal			
4405	5	0	02:30
4941	5	2	02:49
8054	4	2	02:25
12750	5	1	03:43
13690	5	0	01:04
14368	5	0	02:30
16009	5	0	01:14
16291	5	0	02:07
16412	5	0	02:00
16436	5	0	07:58
21945	5	0	01:52
21947	5	0	03:06
21948	5	0	03:43
21964	5	0	05:04
21978	5	0	03:00

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
22002	4	0	06:39
22019	4	1	03:41
22020	4	1	01:34
22037	4	1	01:32
22054	4	1	02:25
22088	4	0	08:16
22106	4	2	06:22
22115	4	1	10:34
22145	4	1	04:32
26421	4	2	02:57
Bird 24: Ovenbird			
3288	5	0	01:27
3363	5	0	01:03
4625	5	0	02:05
8913	5	0	01:12
10778	5	0	00:47
12696	5	0	01:38
12704	4	1	02:36
12715	4	1	02:19
12719	5	0	02:07
12720	4	1	02:06
13861	5	1	01:21
16024	5	1	02:18
16543	5	1	01:17
16898	4	2	01:51
17623	1	5	02:32
17929	1	5	01:27
29778	5	1	04:27
30107	5	1	03:38
30370	5	1	07:06
30372	5	0	04:03
Bird 25: RoseBreastedGrosbeak			
470	3	1	02:27
472	4	2	01:31
978	5	2	03:58
2791	5	1	02:11
3373	5	0	00:45
3394	5	0	02:25
3965	5	0	02:09
3990	5	0	03:00
3996	5	0	01:52
6837	5	0	01:07
6914	5	0	03:52

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
7414	5	0	01:49
8261	5	1	05:16
8299	5	0	02:04
9422	5	1	02:31
11140	5	1	01:34
11173	5	0	01:53
12949	5	0	01:11
14279	3	2	03:41
14354	5	0	02:10
14399	4	2	02:47
15830	5	0	01:24
15887	5	0	02:33
16493	5	1	02:07
Bird 26: ScarletTanager			
949	5	1	01:26
1012	5	0	01:25
1108	4	1	01:31
1363	5	0	01:12
1367	5	0	01:13
1372	5	0	00:55
1932	5	0	02:35
2220	5	0	01:20
3447	5	0	02:41
5756	5	0	03:39
5802	5	1	02:22
5976	5	0	02:21
6262	5	0	01:37
6399	5	0	01:42
6401	5	0	02:08
8306	5	0	01:20
8907	4	0	02:05
9419	5	0	02:18
9435	5	0	02:37
10478	5	1	02:02
11194	4	2	02:51
12255	5	0	02:10
12911	5	0	01:12
13913	5	0	01:13
15681	5	0	00:57
16358	5	0	02:02
Bird 27: SummerTanager			
1020	5	0	03:10
1081	5	0	01:49

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
4045	5	0	02:30
4048	5	0	03:36
4090	5	0	00:31
5657	5	1	02:05
5712	5	0	01:34
7043	5	1	02:21
7049	5	0	02:50
8364	5	0	03:56
9010	5	0	02:38
9914	5	0	03:15
10513	5	0	01:08
13533	5	0	01:36
14695	5	0	01:04
14829	5	0	01:57
15646	5	0	02:50
16562	5	0	02:19
25966	5	0	01:32
28863	5	1	02:02
30810	5	0	01:45
Bird 28: SwampSparrow			
3563	5	0	02:29
4205	5	1	01:47
5076	5	0	02:18
5078	5	2	01:16
5082	5	0	02:10
6053	5	2	02:36
6525	5	1	03:17
6807	4	3	01:36
6937	4	2	02:48
6947	5	0	01:42
7510	5	2	01:36
10923	5	1	01:21
12158	5	2	01:17
13151	5	1	00:55
13474	4	1	02:31
13489	5	0	01:00
14653	5	0	01:28
14657	5	0	01:47
14785	5	0	00:53
14869	5	0	01:45
14871	5	0	00:24
14873	5	0	01:36
14874	5	1	00:45

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
14876	5	0	02:50
14878	5	0	02:07
16485	4	3	00:58
Bird 29: VesperSparrow			
1663	5	1	01:59
1745	5	0	00:38
1768	5	0	01:41
1987	4	1	02:20
2528	5	0	00:26
2596	5	0	00:16
2772	5	1	01:02
2785	5	0	00:41
3201	5	0	01:37
3222	5	0	01:44
3232	5	0	01:13
3233	5	0	01:41
3430	5	1	01:41
4042	5	0	01:46
4086	5	1	01:12
4108	5	1	00:58
4110	5	0	01:02
4532	5	1	02:25
5008	5	0	00:35
5794	5	0	02:18
6403	5	0	01:59
6595	5	0	00:50
6618	5	0	00:15
6782	5	2	02:19
7272	4	4	01:05
7436	5	0	01:25
7485	5	0	02:36
7823	5	1	02:36
7824	5	1	01:38
7875	5	0	04:56
7876	5	0	00:26
7903	5	1	00:58
7907	5	1	01:38
8104	5	1	01:00
8265	5	2	01:58
8528	5	0	00:46
8559	5	0	00:29
8667	5	0	00:56
9140	5	0	00:42

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
10772	5	0	02:56
11405	5	0	00:05
11492	3	5	01:09
12392	5	0	01:20
29567	5	2	02:07
Bird 30: YellowWarbler			
456	5	1	04:25
2531	4	0	02:24
3304	5	2	02:01
3484	5	0	03:06
3722	4	1	02:38
3973	5	1	02:11
4733	5	2	01:13
5063	5	1	02:01
5203	5	1	02:07
5842	5	1	02:09
6051	5	0	01:55
7140	5	0	02:24
8937	5	0	01:36
12119	5	1	00:51
12474	5	0	01:34
13096	5	0	02:24
13475	5	0	00:29
15501	5	0	01:39
15516	5	0	01:57
28480	5	1	03:32
28842	5	0	03:58
Bird 31: ProthonotaryWarbler			
1068	5	0	03:31
1444	5	0	04:54
5695	5	0	02:51
6944	5	1	03:22
7581	5	0	02:33
8419	5	0	02:28
8971	5	0	02:00
8988	5	0	01:24
11834	5	0	01:25
14703	5	0	01:25
14711	5	0	02:15
28971	3	3	03:59
Bird 32: OliveSidedFlycatcher			
1494	5	0	06:01
1561	5	0	00:26

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
1578	5	0	02:57
2046	5	1	01:41
2199	3	0	01:37
2234	4	1	00:54
4363	5	0	01:31
5234	4	2	02:15
5407	5	1	05:15
8675	5	0	00:59
25810	4	1	05:14
29055	5	0	05:04
33365	4	2	03:56
Bird 33: MagnoliaWarbler			
3490	5	0	02:14
3594	5	0	02:39
3991	5	0	01:42
4152	5	0	01:55
4157	5	0	02:12
4295	5	0	02:06
4746	5	0	01:06
4804	4	2	00:43
4844	4	3	00:48
5224	5	0	01:45
5301	5	0	01:15
5317	3	4	00:31
6096	5	0	01:24
6362	5	0	01:40
6373	5	0	00:11
6859	5	0	01:49
8902	5	0	03:15
8921	5	0	01:18
9023	5	0	00:29
9429	3	4	00:52
9442	5	0	01:42
10555	5	1	01:38
11292	5	0	03:26
12869	5	1	00:35
12913	5	0	00:32
14788	5	0	01:02
15111	5	0	00:58
15123	5	0	01:11
15650	5	0	02:32
15816	5	1	01:16
16003	4	2	00:29

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
16224	5	1	01:35
Bird 34: KirtlandsWarbler			
2753	5	0	00:57
2755	5	0	02:21
2758	5	0	00:41
2761	5	0	03:01
2762	5	0	03:46
2764	5	0	02:12
2765	5	0	02:14
2769	5	0	00:48
2771	5	0	01:25
2774	5	0	01:13
2778	5	0	01:21
2779	5	0	01:06
2780	5	0	01:06
2784	5	0	00:50
2787	5	0	01:36
2790	4	1	09:08
2804	5	0	01:27
8585	4	2	04:52
Bird 35: KentuckyWarbler			
7432	5	1	02:05
8367	5	1	02:05
8955	5	0	02:51
9915	5	0	02:25
9932	5	0	01:57
10060	4	1	02:11
10530	5	1	00:38
10655	5	1	01:25
11792	5	0	00:54
12147	5	0	00:42
13112	5	1	01:09
13862	5	0	01:59
14864	4	2	01:55
15468	5	2	01:53
15994	5	2	04:43
16213	5	1	01:22
16252	5	1	01:22
16325	4	3	01:22
16518	4	2	01:33
Bird 36: AmericanGoldfinch			
1782	5	0	02:25
2503	4	2	01:38

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
2595	5	0	01:19
3645	5	1	03:06
5084	3	0	00:47
5909	5	0	01:08
6273	4	2	01:44
6896	5	1	01:39
6997	5	0	01:25
7846	4	1	00:30
8194	4	2	02:19
8860	5	1	01:28
8962	5	0	01:07
9443	5	0	01:02
10044	5	2	00:38
10580	5	1	01:07
11745	5	0	00:54
12060	3	3	00:54
12068	5	1	00:53
12096	4	1	00:39
12101	4	2	00:57
12114	5	3	01:13
12804	5	2	00:51
13428	5	0	01:21
13447	5	1	04:13
13925	5	0	00:47
14272	5	1	00:38
15049	5	1	01:27
15993	5	0	00:48
16437	5	0	01:57
24580	1	5	05:06
33856	5	1	00:14
33858	1	4	00:18
34553	2	4	00:46
Bird 37: AmericanRedstart			
475	5	1	01:24
1078	4	0	02:19
1156	5	1	02:28
2041	5	1	01:57
2504	4	1	01:21
3459	5	1	01:44
3517	5	1	01:02
3521	5	0	02:59
3610	5	1	01:15
4154	5	0	01:39

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
4597	5	3	02:55
4737	5	0	01:19
5734	5	0	01:11
5983	5	1	02:16
6366	4	0	01:01
6853	5	1	02:40
9058	5	1	01:37
9106	5	2	01:03
9109	5	0	01:00
9572	5	0	01:51
10658	5	0	01:15
10935	5	0	00:41
11375	5	0	02:46
12168	5	1	01:14
28794	3	3	01:02
Bird 38: CarolinaChickadee			
6205	5	0	02:05
9026	5	0	01:47
13627	5	1	00:40
17023	5	1	01:17
17029	5	0	02:01
17084	5	2	12:33
17441	4	2	01:18
17566	4	1	05:04
17567	3	2	06:19
Bird 39: BlueGrosbeak			
2524	5	0	05:12
2575	5	2	03:15
3369	5	2	05:50
3436	5	1	02:08
6333	5	1	02:19
7087	5	0	02:03
7166	5	0	01:39
9727	5	0	02:06
10147	5	0	01:23
10187	5	0	00:24
12349	5	0	02:15
14087	5	0	02:40
17118	4	1	02:58
Bird 40: WilsonsWarbler			
3224	5	1	01:30
5150	5	0	02:19
5185	5	0	03:37

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
5190	4	1	03:05
5328	5	0	02:09
7708	4	0	00:44
7758	5	0	01:45
7790	5	0	01:21
7856	5	0	01:27
7867	5	0	01:25
8973	5	1	02:01
8995	5	2	00:37
9484	5	2	01:29
9486	5	2	01:02
11288	5	1	03:02
12221	5	2	01:42
15521	5	1	03:13
15705	5	1	01:08
18826	5	0	00:35
28929	5	0	04:01
28987	5	0	03:32
Bird 41:WhiteEyedVireo			
16283	4	2	05:59
16321	3	2	05:07
16349	3	3	08:05
16373	3	0	10:01
16391	4	1	02:14
16491	4	1	09:57
16528	4	2	05:52
16541	5	1	07:56
16664	5	0	02:25
16878	5	2	03:52
Bird 42: WarblingVireo			
1059	5	0	01:00
1376	5	0	01:33
6856	5	0	01:56
6909	4	2	01:32
7042	5	0	01:42
7772	5	1	01:41
8388	3	1	03:25
10511	5	1	01:19
10928	5	0	00:49
11082	5	0	01:39
13492	5	0	02:32
16882	4	2	03:01
18752	4	2	01:59

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
28458	5	0	03:26
29653	5	0	02:55
Bird 43: SavannahSparrow			
469	3	1	07:35
2129	4	1	02:52
3530	5	0	01:11
4112	5	0	02:00
4114	5	1	02:47
8530	5	1	03:53
11388	5	1	00:54
12436	5	1	02:38
21203	4	2	01:07
23917	5	0	04:36
29705	5	0	03:31
29788	5	0	01:02
Bird 44: NorthernYellowShaftedFlicker			
730	4	2	03:43
1697	5	1	01:09
2342	5	0	00:33
3090	5	0	00:32
3172	5	1	01:32
3186	5	0	01:58
3188	5	0	00:46
3189	5	0	01:03
3242	5	0	01:24
3266	5	0	00:16
4358	5	0	00:21
4579	5	0	00:59
4694	5	0	00:46
4833	5	0	00:43
5294	5	0	00:27
5824	5	0	01:00
5825	5	0	02:49
6695	5	0	00:40
6697	5	0	00:30
6756	5	2	02:11
6760	5	0	02:42
7370	5	0	01:08
8045	5	0	01:05
8743	5	0	00:25
8761	5	0	00:53
9867	5	1	01:41
10836	5	1	01:37

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
Bird 45: FieldSparrow			
670	4	1	02:14
1240	5	1	01:38
1651	5	0	01:33
1776	5	0	02:44
3168	5	1	01:52
3794	5	1	01:26
3827	5	0	01:31
3911	5	0	01:36
6742	4	1	02:12
6744	5	2	02:37
10430	4	0	00:28
11002	5	0	01:03
11006	5	0	00:59
11027	5	0	01:17
24168	5	1	01:30
24242	4	3	02:10
24248	4	1	02:27
32183	4	4	01:18
32202	5	2	01:42
Bird 46: SlateColoredJuncoDarkEyed			
1140	5	0	03:12
1151	5	0	01:08
1274	5	0	01:28
2763	5	0	01:33
3567	5	0	01:01
3567	5	1	00:50
3589	5	0	00:03
3638	5	0	00:09
3646	5	0	00:24
3650	5	0	02:17
3657	5	0	01:38
3661	5	0	01:06
4164	5	0	00:44
4266	5	0	01:52
4767	5	0	01:35
4805	5	0	00:46
4866	5	0	01:39
5345	5	0	02:00
5454	5	0	00:30
7288	5	0	01:11
7958	4	2	01:53
8040	4	2	00:52

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
8069	5	0	01:02
10391	5	2	00:53
21129	5	1	00:59
21278	3	3	00:51
24092	5	2	00:41
Bird 47: WillowFlycatcher			
5823	5	1	02:22
6384	3	4	02:56
7743	5	0	00:57
8580	5	0	01:38
10064	5	0	02:02
10647	5	0	01:53
12399	5	0	01:32
29018	5	1	01:50
29313	2	4	04:07
29748	5	1	08:39
29749	5	0	01:53
Bird 48: WinterNorthernWren			
601	4	4	07:11
2040	5	0	02:57
2073	3	1	04:08
2955	3	3	02:54
4319	5	2	08:23
4802	4	1	03:31
5431	5	2	04:02
10880	5	3	03:56
21598	5	1	05:06
Bird 49: WesternMeadowlark			
441	5	1	06:21
4696	4	3	08:14
5774	5	2	08:57
5816	5	1	08:30
6610	5	0	03:09
7395	5	2	04:44
7416	5	0	06:05
8769	5	3	11:01
Bird 50: YellowThroatedWarbler			
909	5	0	01:07
1122	5	0	01:20
2330	3	1	05:00
3127	5	0	04:54
4438	5	0	01:52
4488	5	0	04:11

File Number	Recording Quality	Background Noise Level	File Length (mm:ss)
4511	5	0	02:12
4960	5	1	02:13
5677	5	0	01:32
6302	5	1	01:14
7620	5	1	01:51
10532	5	1	01:40
14308	5	0	02:26
16212	4	3	01:42

List of References

[1] Virkkala R, Lehikoinen A. Patterns of climate-induced density shifts of species: poleward shifts faster in northern boreal birds than in southern birds. Global change biology. 2014;20(10):2995–3003.

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

[3] Jančovič P, Köküer M. Automatic detection and recognition of tonal bird sounds in noisy environments. EURASIP Journal on Advances in Signal Processing. 2011;2011(1):982936.

[4] Briggs F, Lakshminarayanan B, Neal L, Fern XZ, Raich R, Hadley SJ, et al. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. The Journal of the Acoustical Society of America. 2012;131(6):4640–4650.

[5] Jančovič P, Köküer M. Estimation of voicing-character of speech spectra based on spectral shape. Signal Processing Letters, IEEE. 2007;14(1):66–69.

[6] Jančovič P, Köküer M. 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. IEEE; 2011. p. 517–520.

[7] Kogan JA, Margoliash D. 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. 1998;103(4):2185–2196.

[8] Anderson SE, Dave AS, Margoliash D. Template-based automatic recognition of birdsong syllables from continuous recordings. The Journal of the Acoustical Society of America. 1996;100(2):1209–1219.

[9] Chen Z, Maher RC. Semi-automatic classification of bird vocalizations using spectral peak tracks. The Journal of the Acoustical Society of America. 2006;120(5):2974–2984.

[10] McIlraith A, Card H. Bird song identification using artificial neural networks and statistical analysis. In: Electrical and Computer Engineering, 1997. Engineering Innovation: Voyage of Discovery. IEEE 1997 Canadian Conference on. vol. 1. IEEE; 1997. p. 63–66.

[11] Härmä A, Somervuo P. Classification of the harmonic structure in bird vocalization. In: Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on. vol. 5. IEEE; 2004. p. V–701.

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

[13] Heller JR, Pinezich JD. Automatic recognition of harmonic bird sounds using a frequency track extraction algorithm. The Journal of the Acoustical Society of America. 2008;124(3):1830–1837.

[14] Cheng J, Sun Y, Ji L. A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. Pattern Recognition. 2010;43(11):3846–3852.

[15] Picone JW. Signal modeling techniques in speech recognition. Proceedings of the IEEE. 1993;81(9):1215–1247.

[16] Campbell Jr JP. Speaker recognition: a tutorial. Proceedings of the IEEE. 1997;85(9):1437–1462.

[17] Rabiner L, Juang BH. Fundamentals of speech recognition. 1993;.

[18] Furui S. Digital Speech Processing, Synthesis, and Recognition (Revised and Expanded). Digital Speech Processing, Synthesis, and Recognition (Second Edition, Revised and Expanded). 2000;.

[19] Alpaydin E. Introduction to machine learning (adaptive computation and machine learning series). The MIT Press Cambridge; 2004.

[20] Furui S. Recent advances in speaker recognition. In: Audio-and Video-based Biometric Person Authentication. Springer; 1997. p. 235–252.

[21] Chen K, Wang L, Chi H. Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification. International Journal of Pattern Recognition and Artificial Intelligence. 1997;11(03):417–445.

[22] Quatieri TF. Discrete-time speech signal processing: principles and practice. Pearson Education India; 2002.

[23] Reynolds DA, Rose RC. Robust text-independent speaker identification using Gaussian mixture speaker models. Speech and Audio Processing, IEEE Transactions on. 1995;3(1):72–83.

[24] Mashao DJ, Skosan M. Combining classifier decisions for robust speaker identification. Pattern Recognition. 2006;39(1):147–155.

[25] Huang X, Acero A, Hon HW, et al.. Spoken language processing. Prentice Hall Englewood Cliffs; 2001.

[26] Kwan C, Ho K, Mei G, Li Y, Ren Z, Xu R, et al. An automated acoustic system to monitor and classify birds. EURASIP Journal on Advances in Signal Processing. 2006;2006(1):1–19.

[27] Hong Q, Kwong S. A discriminative training approach for text-independent speaker recognition. Signal Processing. 2005;85(7):1449–1463.

[28] Reynolds D. Gaussian mixture models. Encyclopedia of Biometrics. 2015;p. 827–832.

[29] Clemins PJ. Automatic classification of animal vocalizations. Faculty of the Graduate School, Marquette University; 2005.

[30] Gish H, Schmidt M. Text-independent speaker identification. Signal Processing Magazine, IEEE. 1994;11(4):18–32.

[31] McLachlan GJ, Basford KE. Mixture models. New York: Marcel Dekker; 1988.

[32] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological). 1977;p. 1–38.

[33] Bishop CM. Neural networks for pattern recognition. Oxford university press; 1995.

[34] Stuttle MN. A Gaussian mixture model spectral representation for speech recognition. University of Cambridge; 2003.

[35] Trentin E, Gori M. A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing. 2001;37(1):91–126.

[36] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989;77(2):257–286.

[37] Fink GA. Markov Models for Pattern Recognition: From Theory to Applications. Advances in Computer Vision and Pattern Recognition. Springer London; 2014. Available from: `https://books.google.co.uk/books?id=3c-4BAAAQBAJ`.

[38] Young S, Evermann G, Kershaw D, Moore G, Odell J, Ollason D, et al. The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department. Cambridge, UK, December. 2002;.

[39] Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The annals of mathematical statistics. 1970;41(1):164–171.

[40] Bilmes JA, et al. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute. 1998;4(510):126.

[41] Maimon O, Rokach L. Data mining and knowledge discovery handbook. vol. 2. Springer; 2005.

[42] Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. The computer journal. 1998;41(8):578–588.

[43] Castro V, Yang J. A fast and robust general purpose clustering algorithm. In: Fourth European Workshop on Principles of Knowledge Discovery in Databases and Data Mining; 2000. .

[44] Sneath P, Sokal R. Numerical Taxonomy WH Freeman & Co San Francisco. USA; 1973.

[45] Everitt BS, Landau S, Leese M, Stahl D. Hierarchical clustering. Cluster Analysis, 5th Edition. 2011;p. 71–110.

[46] Härmä A. Automatic identification of bird species based on sinusoidal modeling of syllables. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on. vol. 5. IEEE; 2003. p. V–545.

[47] Selin A, Turunen J, Tanttu JT. Wavelets in recognition of bird sounds. EURASIP Journal on Applied Signal Processing. 2007;2007(1):141–141.

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

[49] McAulay RJ, Quatieri TF. Speech analysis/synthesis based on a sinusoidal representation. Acoustics, Speech and Signal Processing, IEEE Transactions on. 1986;34(4):744–754.

[50] Lee CH, Lee YK, Huang RZ. Automatic recognition of bird songs using cepstral coefficients. Journal of Information Technology and Applications. 2006;1(1):17–23.

[51] Chou CH, Liu PH, Cai B. On the studies of syllable segmentation and improving MFCCs for automatic birdsong recognition. In: Asia-Pacific Services Computing Conference, 2008. APSCC'08. IEEE. IEEE; 2008. p. 745–750.

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

[53] Jancovic P, Kokuer M, Russell M. 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. IEEE; 2014. p. 8252–8256.

[54] McIlraith AL, Card HC. Birdsong recognition using backpropagation and multivariate statistics. Signal Processing, IEEE Transactions on. 1997;45(11):2740–2748.

[55] Catchpole CK, Slater PJ. Bird song: biological themes and variations, 2nd Edition. Cambridge university press; 2008.

[56] Oberweger K, Goller F. The metabolic cost of birdsong production. Journal of Experimental Biology. 2001;204(19):3379–3388.

[57] Ward S, Lampe HM, Slater PJ. Singing is not energetically demanding for pied flycatchers, Ficedula hypoleuca. Behavioral Ecology. 2004;15(3):477–484.

[58] Alderleaf Wilderness College - Bird Communication; 2015. Accessed: 27/12/15. `http://www.wildernesscollege.com/bird-communication.html`.

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

[60] Marler PR, Slabbekoorn H. Nature's music: the science of birdsong. Academic Press; 2004.

[61] White G. The natural history and antiquities of Selborne. Methuen; 1901.

[62] Howard HE. Territory in bird life. J. Murray; 1920.

[63] Krebs J. Habituation and song repertoires in the great tit. Behavioral Ecology and Sociobiology. 1976;1(2):215–227.

[64] Falls JB. Does song deter territorial intrusion in White-throated Sparrows (Zonotrichia albicollis)? Canadian Journal of Zoology. 1988;66(1):206–211.

[65] Nowicki S, Peters S, Podos J. Song learning, early nutrition and sexual selection in songbirds. American Zoologist. 1998;38(1):179–190.

[66] Eriksson D, Wallin L. Male bird song attracts females-a field experiment. Behavioral Ecology and Sociobiology. 1986;19(4):297–299.

[67] Mountjoy DJ, Lemon RE. Female choice for complex song in the European starling: a field experiment. Behavioral Ecology and Sociobiology. 1996;38(1):65–71.

[68] Johnson LS, Searcy WA. Female attraction to male song in house wrens (Troglodytes aedon). Behaviour. 1996;133(5):357–366.

[69] Searcy WA. Measuring responses of female birds to male song. In: Playback and studies of animal communication. Springer; 1992. p. 175–189.

[70] Clayton N. Song discrimination learning in zebra finches. Animal Behaviour. 1988;36(4):1016–1024.

[71] Langmore NE. Functions of duet and solo songs of female birds. Trends in Ecology & Evolution. 1998;13(4):136–140.

[72] Cooney R, Cockburn A. Territorial defence is the major function of female song in the superb fairy-wren, Malurus cyaneus. Animal Behaviour. 1995;49(6):1635–1647.

[73] Baptista LF, Trail PW, DeWolfe BB, Morton ML. Singing and its functions in female white-crowned sparrows. Animal Behaviour. 1993;46(3):511–524.

[74] Collins SA. Is female preference for male repertoires due to sensory bias? Proceedings of the Royal Society of London B: Biological Sciences. 1999;266(1435):2309–2314.

[75] Lack D, Gillmor R. The life of the robin. Witherby London; 1965.

[76] Trainer JM, McDonald DB, Learn WA. The development of coordinated singing in cooperatively displaying long-tailed manakins. Behavioral Ecology. 2002;13(1):65–69.

[77] HOELZEL A. Song characteristics and response to playback of male and female robins Erithacus rubecula. Ibis. 1986;128(1):115–127.

[78] Yasukawa K, Searcy WA. Song repertoires and density assessment in red-winged blackbirds: further tests of the Beau Geste hypothesis. Behavioral Ecology and Sociobiology. 1985;16(2):171–175.

[79] Garamszegi LZ, Pavlova DZ, Eens M, Møller AP. The evolution of song in female birds in Europe. Behavioral Ecology. 2007;18(1):86–96.

[80] Armstrong EA. Discovering bird song. Shire Publications; 1975.

[81] Kacelnik A, Krebs JR. The dawn chorus in the great tit (Parus major): proximate and ultimate causes. Behaviour. 1983;83(3):287–308.

[82] Thomas RJ, Széskely T, Cuthill IC, Harper DG, Newson SE, Frayling TD, et al. Eye size in birds and the timing of song at dawn. Proceedings of the Royal Society of London B: Biological Sciences. 2002;269(1493):831–837.

[83] Berg KS, Brumfield RT, Apanius V. Phylogenetic and ecological determinants of the neotropical dawn chorus. Proceedings of the Royal Society of London B: Biological Sciences. 2006;273(1589):999–1005.

[84] Morton E. A comparison of vocal behavior among tropical and temperate passerine birds. Ecology and evolution of acoustic communication in birds. 1996;p. 258–268.

[85] Göth A, Vogel U, Curio E. The acoustic communication of the Polynesian megapode Megapodius pritchardii GR Gray. Zoologische Verhandelingen. 1999;p. 37–52.

[86] LEVIN RN. Song behaviour and reproductive strategies in a duetting wren, Thryothorus nigricapillus: I. Removal experiments. Animal Behaviour. 1996;52(6):1093–1106.

[87] LEVIN RN. Song behaviour and reproductive strategies in a duetting wren, Thryothorus nigricapillus: II. Playback experiments. Animal Behaviour. 1996;52(6):1107–1117.

[88] Borror DJ. Intraspecific variation in passerine bird songs. The Wilson Bulletin. 1961;p. 57–78.

[89] Collias N, Joos M. The spectrographic analysis of sound signals of the domestic fowl. Behaviour. 1953;5(1):175–188.

[90] Wolff D. Detecting bird sounds via periodic structures: a robust pattern recognition approach to unsupervised animal monitoring. To be found at< http://www-mmdb iai uni-bonn de/download/Diplomarbeiten/Diplomarbeit_Daniel_Wolff pdf>[quoted 0106 2011]. 2008;.

[91] Stowell D, Plumbley MD. Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers. Centre for Digital Music, Queen Mary, University of London, London, UK, Tech Rep C4DM-TR-09-12. 2011;181.

[92] Briefer E, Osiejuk TS, Rybak F, Aubin T. Are bird song complexity and song sharing shaped by habitat structure? An information theory and statistical approach. Journal of Theoretical Biology. 2010;262(1):151–164.

[93] EcoBirds; 2015. Accessed: 27/12/15. `http://www.birds.ecoport.org/Behaviour/EBbirdsong.htm`.

[94] Wild JM. Neural pathways for the control of birdsong production. Journal of neurobiology. 1997;33(5):653–670.

[95] Marler P. Tonal quality of bird sounds. Bird vocalizations. 1969;p. 5–18.

[96] Nowicki S, Marler P. How do birds sing? Music Perception. 1988;p. 391–426.

[97] Nowicki S, Marler P, Maynard A, Peters S. Is the tonal quality of birdsong learned? Evidence from song sparrows. Ethology. 1992;90(3):225–235.

[98] Doupe AJ, Kuhl PK. Birdsong and human speech: common themes and mechanisms. Annual review of neuroscience. 1999;22(1):567–631.

[99] Baptista L, Kroodsma D. Avian bioacoustics. Handbook of the birds of the world. 2001;6:11–52.

[100] Fant G. Acoustic theory of speech production.'s-Gravenhage: Mouton and Co. I960. 1960;.

[101] Westneat MW, Long J, Hoese W, Nowicki S. Kinematics of birdsong: functional correlation of cranial movements and acoustic features in sparrows. The Journal of experimental biology. 1993;182(1):147–171.

[102] Hoese WJ, Podos J, Boetticher NC, Nowicki S. Vocal tract function in birdsong production: experimental manipulation of beak movements. Journal of Experimental Biology. 2000;203(12):1845–1855.

[103] Podos J, Sherer JK, Peters S, Nowicki S. Ontogeny of vocal tract movements during song production in song sparrows. Animal Behaviour. 1995;50(5):1287–1296.

[104] Podos J. A performance constraint on the evolution of trilled vocalizations in a songbird family (Passeriformes: Emberizidae). Evolution. 1997;p. 537–551.

[105] Pieplow N, DiGiorgio M, Peterson RT. Peterson Field Guide to Bird Sounds of Eastern North America. Houghton Mifflin Harcourt; 2017.

[106] Oden AI. Changes in Avian Vocalization Occurrence and Frequency Range During the Winter. 2013;.

[107] Thorpe WH. The learning of song patterns by birds, with especial reference to the song of the chaffinch Fringilla coelebs. Ibis. 1958;100(4):535–570.

[108] Poulsen H. Inheritance and learning in the song of the chaffinch (Fringilla coelebs L.). Behaviour. 1951;3(1):216–242.

[109] Poulsen H. The calls of the Chaffinch (Fringilla coelebs L.) in Denmark. Dan Ornithol Foren Tidsskr. 1958;52:89–105.

[110] Marler P. A comparative approach to vocal learning: song development in White-crowned Sparrows. Journal of comparative and physiological psychology. 1970;71(2p2):1.

[111] Wada H. The development of birdsong. Nature Education Knowledge. 2012;3(10):86.

[112] Beecher M, George F, Michel Le M RF. Birdsong and Vocal Learning during Development. Encyclopedia of Behavioral Neuroscience. Oxford: Academic Press; 2010.

[113] Brenowitz EA, Margoliash D, Nordeen KW. An introduction to birdsong and the avian song system. Journal of neurobiology. 1997;33(5):495–500.

[114] Brainard MS, Doupe AJ. What songbirds teach us about learning. Nature. 2002;417(6886):351–358.

[115] Konishi M. Effects of Deafening on Song Development in American Robins and Black-headed Grosbeaks3. Zeitschrift für Tierpsychologie. 1965;22(5):584–599.

[116] Bottjer SW, Johnson F. Circuits, hormones, and learning: vocal behavior in songbirds. Journal of neurobiology. 1997;33(5):602–618.

[117] Schlinger BA. Sex steroids and their actions on the birdsong system. Journal of neurobiology. 1997;33(5):619–631.

[118] Stowell D, Plumbley MD. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ. 2014;2:e488.

[119] xeno-canto; 2015. Accessed: 17/4/15. `https://www.xeno-canto.org`.

[120] Borror Laboratory of Bioacoustics;. The Ohio State University, Columbus, OH, all rights reserved. `www.blb.osu.edu`.

[121] Macaulay Library, Cornell Laboratory of Ornithology; 2015. Accessed: 17/4/15. `https://www.macaulaylibrary.org`.

[122] Cornell Lab of Ornithology 159 Sapsucker Woods Rd Ithaca, NY 14850; 2015. Accessed: 17/4/15. `https://www.birds.cornell.edu`.

[123] ICML int. Conf. Proc. 1st workshop on Machine Learning for Bioacoustics - ICML4B; 2013. Http://sabiod.univ-tln.fr.

[124] Briggs F, Raich R, Eftaxias K, Lei Z, Huang Y. The ninth annual MLSP competition: overview. In: IEEE International workshop on machine learning for signal processing, Southampton, United Kingdom., Sept; 2013. p. 22–25.

[125] NIPS Int. Conf. Proc. Neural Information Processing Scaled for Bioacoustics, from Neurons to Big Data; 2013. Http://sabiod.org/nips4b.

[126] Goëau H, Glotin H, Vellinga WP, Planqué R, Rauber A, Joly A. LifeCLEF bird identification task 2014. In: CLEF2014; 2014. .

[127] The ICML 2013 Bird Challenge; 2015. Accessed: 20/5/15. https://www.kaggle.com/c/the-icml-2013-bird-challenge.

[128] Museum national d'Histoire naturelle; 2015. Accessed: 20/5/15. https://www.mnhn.fr/fr.

[129] MLSP 2013 Bird Classification Challenge; 2015. Accessed: 20/5/15. https://www.kaggle.com/c/mlsp-2013-birds.

[130] Biotope, The Ecology Consultancy Firm; 2015. Accessed: 20/5/15. http://www.biotope.fr/en.

[131] Multi-label Bird Species Classification - NIPS 2013; 2015. Accessed: 20/5/15. https://www.kaggle.com/c/multilabel-bird-species-classification-nips2013.

[132] Transcriber - Copyright (C) 1998-2008, DGA; 2015. Accessed: 17/4/15. http://trans.sourceforge.net.

[133] Jancovic P, Kokuer M, Zakeri M, Russell M. Unsupervised discovery of acoustic patterns in bird vocalisations employing DTW and clustering. In: Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European. IEEE; 2013. p. 1–5.

[134] Jancovic P, Zakeri M, Kokuer M, Russell M. HMM-based modelling of individual syllables for bird species recognition from audio field recordings. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE; 2015. p. 768–772.

[135] Smith JOIII. Spectral Audio Signal Processing. W3K; 2011.

[136] Miller HJ, Han J. Geographic data mining and knowledge discovery. CRC Press; 2009.

[137] Basu S, Davidson I, Wagstaff K. Constrained clustering: Advances in algorithms, theory, and applications. CRC Press; 2008.

[138] Jancovic P, Kokuer M, Murtagh F. Reliability-based estimation of the number of noisy features: Application to model-order selection in the union models. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on. vol. 1. IEEE; 2003. p. I–416.

[139] Young S, Kershaw D, Odell J, Ollason D, Valtchev V, Woodland P. The HTK Book V2. 2. Entropic Ltd., Jan; 1999.

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

[141] Zhang X, Li Y. Adaptive energy detection for bird sound detection in complex environments. Neurocomputing. 2015;155:108–116.

[142] Evangelista TL, Priolli TM, Silla Jr CN, Angelico BA, Kaestner CA. Automatic segmentation of audio signals for bird species identification. In: Multimedia (ISM), 2014 IEEE International Symposium on. IEEE; 2014. p. 223–228.

[143] Ventura TM, de Oliveira AG, Ganchev TD, de Figueiredo JM, Jahn O, Marques MI, et al. Audio parameterization with robust frame selection for improved bird identification. Expert Systems with Applications. 2015;42(22):8463–8471.

[144] Fagerlund S, Laine UK. New parametric representations of bird sounds for automatic classification. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE; 2014. p. 8247–8251.

[145] Tyagi H, Hegde RM, Murthy HA, Prabhakar A. Automatic identification of bird calls using spectral ensemble average voice prints. In: Signal Processing Conference, 2006 14th European. IEEE; 2006. p. 1–5.

[146] Ptacek L, Machlica L, Linhart P, Jaska P, Muller L. Automatic recognition of bird individuals on an open set using as-is recordings. Bioacoustics. 2016;25(1):55–73.

[147] Fagerlund S. Automatic recognition of bird species by their sounds. Finlandia: Helsinki University Of Technology. 2004;.