Quantifying difference in vocalizations of bird populations

Colm O’Reilly¹, Nicola M. Marples², David J. Kelly², Naomi Harte¹

¹Sigmaedia, Department of Electrical & Electronic Engineering, Trinity College Dublin, Ireland
²Trinity Centre for Biodiversity Research & Department of Zoology, Trinity College Dublin, Ireland

oreilcl6@tcd.ie, nharte@tcd.ie

Abstract

Populations of a bird species can evolve over time to become a new species. While plumage patterns and other morphological information can remain constant, the vocalizations of a given population may have diversified enough to warrant reclassification. Thus ornithologists are interested in how to measure call and song similarity in birds in a systematic, repeatable manner. Given the success of speech processing methods applied to bird species classification, this paper presents work on developing a measure of bird call similarity. The method is inspired by human speech dialect separation measurement using a representation of the pitch contour micro-structure. The measure is applied to bird populations with calls that are considered very similar, very different and between these two extremes. Initial results are very promising, with the behavior of the metric consistent with accepted levels of similarity for the populations tested to date.

Index Terms: Dialect separation, pitch contour, bird calls

1. Introduction

Speech processing techniques have been successfully exploited in the analysis of calls and songs from bird populations. Work to date has focused on species classification, inspired by automatic speaker recognition, using recordings of calls and song [1, 2, 3, 4, 5, 6, 7, 8]. A common topic is syllable or phrase classification [9, 10, 11, 12] and song structure [13].

A common issue in ornithology is to determine how similar calls or songs between subspecies of a single bird are. This has important implications for the correct taxonomic classification of populations. Work by Harte et al. in [14] was the first to examine the issue of call similarity in this context. That paper demonstrated that classifier performance is related to similarity but is not a quantifiable indicator. Standard approaches in ornithology to birdsong analysis are laborious, subjective and can lack repeatability. McKay et al. in [15] examined song in making a case for the Bahamian subspecies of the Yellow-throated Warbler to be reclassified as a distinct species. Song divergence was important evidence in the reclassification process. Comparisons were on the basis of visual inspection of spectrograms. Sangster et al. in [16], describe a new species of owl as the Rinjani Scops Owl, based on analysis of spectrograms. Harte et al. in [14] extended it to the domain of bird calls and short songs. Thus we report the first stage of work in developing a measure of call similarity for birds to aid ornithologists in their work.

2. Dialect Distance

Pitch contours provide cues for tasks such as language and dialect identification. Lin and Wang in [19] used Legendre polynomial coefficients to represent segments of pitch contours for language identification. Biadsy and Hirschberg [20] demonstrated that four Arabic dialects exhibited significant differences from one another in terms of prosodic structure, including pitch range, register, and pitch dynamics. After human perception tests indicated that prosodic cues, including pitch movements, can be used to identify one accent from another, Mehrabani et al., in [17, 18], explored this in a system for measuring separation between dialects via differences in pitch contour microstructures. Pitch contour analysis does not require prior knowledge, manual segmentation or pre-processing of the speech. Figure 1a shows the pitch values from a speech sample. Each window extracts 3 non-zero fundamental frequency values \( F_0 \) from speech recordings of different dialects, \( D_i \), with an overlap of two samples between adjacent windows. Each window \( F_{01}, F_{02}, F_{03} \), is then transformed into a 2D vector \( \{F_{02} - F_{01}, F_{03} - F_{02}\} \). This 2D vector is used to generate a code book for each dialect being compared, i.e. \( D_1 \) and \( D_2 \). An example of a codebook can be seen in Figure 1b. For human speech analysis, threshold 1 was set at 3Hz. A more in depth description of the dialect distance extraction is available in [17, 18] but a brief summary is provided here to aid in understanding the current work.

\[
P(w_i|w_j) = \frac{C(w_i, w_j)}{C(w_i)} i, j = 1, 2, ..., 9
\]  

After extracting \( F_0 \) and counting the occurrences of each pitch code, the conditional probabilities are then calculated using Eq. 1 for each class and stored in a \([9x9]\) matrix.
Figure 2: Average spectra of (a) BNO BU (top) & WW (bottom), (b) OBSB BU & SU (top) & WW (bottom), (c) AF (top), BS (bottom), (d) Mexican Antbirds: DAB (top), DWA (middle), & MAT (top) data sets.

In Eq. 1 \( C(w_j, w_i) \) is the number of transitions from code \( w_j \) to code \( w_i \), \( P(w_j) \) is the number of occurrences of code \( w_j \), and \( P(w_j | w_i) \) is the probability of transition. Next the Kullback Leibler (KL) divergence was used to derive a distance metric between the different populations. The KL divergence or relative entropy is a non-commutative measure of similarity/dissimilarity between distributions or statistical models [17].

Work in [17] focused on dialect similarity using pitch pattern bi-gram methods. While work in [18] also explored other metrics including log likelihood scores distribution, pitch pattern uni-gram and energy pattern bi-gram methods, only the bi-gram pitch pattern method is used here, which was found to be a robust indicator of dialect separation [18].

3. Measuring bird vocalization similarity

3.1. Bird Data

Four different data sets were used for experiments in this paper. Set 1 contains two subspecies which we anticipated to have similar calls. Set 2 contains two subspecies from three locations. Vocalizations from two locations are from one subspecies. The third population may be split into a separate species following genetic analysis. Differences in the calls of this population will add weight to its reclassification. Set 3 contains two different passerine species from different families, which means they are very distantly related. Finally set 4 contains three sets of recordings from different species of Mexican Antbird, which produce complex song. Each population of Antbirds is a different species. The amount of data in each set is summarized in Table 1. The average spectrum for each set of bird data is shown in Figure 2.

3.1.1. Indonesian Island Populations - Set 1 & 2

Recordings of Black-naped Orioles (BNO) and the Olive-backed Sunbirds (OBSB) calls were collected by Trinity College Dublin (TCD) Zoology Department during field trips to Indonesia. Harte et al. used this data to build Gaussian Mixture Model (GMM) classifiers in [14].

The BNOs found on the Wakatobi islands (including Wangi-Wangi (WW)), most likely colonized from the islands of Sulawesi (SU) and Buton (BU). The BU and WW populations of BNO were investigated for this paper as Set 1. The BU population of BNO is recognized as Oriolus chinensis celebensis. The WW population is a different subspecies recognized as Oriolus chinensis honeratensis. As subspecies of a single species, we predicted their calls would be similar [14].

Three island populations of OBSB are studied here as Set 2, with calls from BU, SU and WW. The OBSBs from mainland SU and BU are officially recognized as subspecies Cinnyris juliarius plateni. The birds found on WW, are officially recognized as subspecies Cinnyris juliarius infrenatus, but there is strong evidence that the Wakatobi birds may be a separate species. Whilst the reclassification of any bird is a multifaceted task, the individuality of their song is as important as morphological and plumage differences [14]. If the WW bird vocalizations show significant and repeatable differences from the BU & SU OBSB populations, this will add significant weight to its classification as a different species.

3.1.2. Species - Alder Flycatcher & Barn Swallow - Set 3

Set 3 contains two species from different families: the Alder Flycatcher (AF) and the Barn Swallow (BS). The species classifier in [6] by Graciarena et al., showed both species to have a low Equal Error Rate (EER). This is consistent with their taxonomic separation. Using an online database "http://www.xeno-canto.org/" [21] we acquired 33 AF recordings and 77 BS recordings labeled as ‘A’ quality. These recordings required background bird vocalization removal and segmentation which was done in MATLAB using a modified version of ‘Harma Syllable Segmentation’ by Michael Lindermuth [22]. After pre-processing there were approximately 500 syllables for each species. The recordings here are thus not from the same dataset used in [6].

3.1.3. Mexican Antbirds - Set 4

Chu et al. used an Antbird corpus in [23], which was also used by Trifa et al. in [1]. These recordings are songs which are more complex than the calls of sets 1-3. The Mexican Antrush is part of the Formicariidae family while the Dot winged Antwren and the Dusky Antbird are part of the Thamnophilidae family. For ease, they will be referred to as MAT, DWA and DAB respectively. We predict the MAT’s vocalizations will be more different than the DWA and DAB’s, but it should still be possible to separate DAB and DWA calls as they are different

Table 1: Bird data for sets 1-4

<table>
<thead>
<tr>
<th>Species</th>
<th>No. of Calls</th>
<th>No. of Birds</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNO BU</td>
<td>110</td>
<td>15</td>
</tr>
<tr>
<td>Wangi-Wangi</td>
<td>77</td>
<td>10</td>
</tr>
<tr>
<td>OBSB BU &amp; SU</td>
<td>517</td>
<td>11</td>
</tr>
<tr>
<td>Sulawesi</td>
<td>218</td>
<td>2</td>
</tr>
<tr>
<td>Wangi-Wangi</td>
<td>608</td>
<td>10</td>
</tr>
<tr>
<td>Species</td>
<td>No. of Syllables</td>
<td>No. of Recordings</td>
</tr>
<tr>
<td>Adler Flycatcher</td>
<td>515</td>
<td>33*</td>
</tr>
<tr>
<td>Dusky Antbird</td>
<td>888</td>
<td>o</td>
</tr>
<tr>
<td>Dot winged Antwren</td>
<td>159</td>
<td>o</td>
</tr>
<tr>
<td>Mexican Antrush</td>
<td>609</td>
<td>o</td>
</tr>
</tbody>
</table>

* Not known if each recording contains a unique bird subject, ○ For the Mexican Antbirds the number of unique birds is not specified but all recordings are from the same location.
Figure 3: Subspecies distance method block diagram [17, 18].

The analysis of these birds is compared with other bird populations later in this paper.

3.2. System Modification

The work presented in [17, 18] showed how pitch contours can be used to derive a dialect similarity measure. This prompted the question can similar ideas be applied to bird calls to measure population similarity? There are fundamental differences between human speech and bird vocalizations. The anatomy of a bird is clearly different to a human. The sound-producing organ for birds, the syrinx, is similar to a human larynx except it is located at the top of the bronchus which allows birds to rapidly change pitch [24]. Bird vocalizations tend to have higher fundamental frequency and are capable of changing pitch much more rapidly. To account for this, modifications needed to be made to the system in [17, 18]. Here instead of dealing with dialect 1 and 2, we are measuring the similarity of species/population 1 (S1) and species/population 2 (S2). An overview of our proposed modified distance system inspired by [17, 18] is given in Figure 3.

The first stage of our system involves extracting the fundamental frequency (\(F_0\)). In [17, 18] RAPT (taken from Voicebox for MATLAB [25]) was used to extract the \(F_0\) as they were dealing with human speech. YIN [26] is more accurate when dealing with higher pitch signals like bird calls. YIN is used by Sound Analysis Pro (SAP) for \(F_0\) extraction and is discussed by Mandelblat-Cerf in [27].

Window sizes of 25, 30, 10 and 10 ms were selected for the BNO, OBSB, Mexican Antbirds and AF & BS respectively. The frame rate was 5 ms for all sets. These values were chosen experimentally by observing the best fit pitch contour line over the spectrograms. Examples of pitch extraction using YIN for each data set can be seen in Figure 4.

These pitch samples are then used to create pitch contours which are mapped to a codebook. Each population has a 3x3 codebook. Figure 5 shows the histograms of the OBSB populations. The different distribution of the BU and WW pitch contours suggests a difference between the two populations.

To establish a threshold, we plotted histograms of \(\Delta F_{01}\) and \(\Delta F_{02}\). Threshold 1 was chosen so that approximately 70% of each histogram was within \pm\ threshold 1 for each species. In [17, 18], Mehrabani et al. used a threshold of \pm3 Hz. After analyzing two hours of male speech, approximately 70% of the \(F_0\) was found to be within \pm3 Hz, hence 70% was used as a target for each species’ threshold as a reasonable first choice. A species specific threshold was established as 20, 50, 50 and 150 Hz for the BNO, OBSB, Antbirds and AF/BS combination, respectively. These simple modifications were used to adapt the original system from [17, 18] to form a new system used here to assess bird call similarity between populations.

4. Results

4.1. Classification Results

To give an indication of how separable the birds calls are, each set of data was used to build a classifier for the corresponding birds. We used 12 MFCCs, log energy, delta and double delta coefficients as our features extracted using a 25 ms window with a frame rate of 10 ms. GMMs were trained with varying numbers of mixtures, dictated by the amount of data available for each set. Results using these models are shown in Table 2. The BNO has a low classification rate which is expected when the classes are the same species. The Antbirds and AF vs. BS classification rates are high because the data sets contain different species. The OBSB classes are correctly recognized 93.82% of the time, even though the populations are currently classed as different subspecies of the same species.

<table>
<thead>
<tr>
<th>Data</th>
<th>Classification Rate (%)</th>
<th>No of Mixtures</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNO</td>
<td>95.67</td>
<td>5</td>
</tr>
<tr>
<td>OBSB</td>
<td>93.82</td>
<td>20</td>
</tr>
<tr>
<td>Antbirds</td>
<td>98.12</td>
<td>256</td>
</tr>
<tr>
<td>AF vs. BS</td>
<td>93.54</td>
<td>20</td>
</tr>
</tbody>
</table>

4.2. Separation measure with 3x3 Codebook

The first experiment uses the modified system with a 3x3 codebook. Table 3 column 1 shows the 3x3 codebook results for Sets 1-4. The two BNO populations yield a small distance of 0.0455 for populations with very similar calls. The distances reported for the MAT compared to both the DWA and DAB yield the largest distances of 0.8156 and 0.822 respectively. This is consistent with the expectation that the song from these birds is very distinct. The DAB and DWA comparison and AF versus BS yield comparable distances of 0.2278 and 0.3588. These measures are smaller, but are still from quite distinct populations. Thus measures of between .35 and .82 all correspond to identifiable difference in call or song. The measure of the OBSB at 0.5155 falls between these extremes. The large distance for the OBSB supports the hypothesis that the subspecies in question has diversified, supporting the suggestion that the WW population be reclassified as a separate species. These initial results are very promising but our measure of similarity still requires refinement. As a first step, we decided to explore whether allowing a more detailed representation of the pitch slopes for a given species was useful. Thus the 3x3 codebook was extended to a 5x5 codebook by introducing...
Figure 5: 3D histogram examples $\Delta F_{01}$ vs. $\Delta F_{02}$ vs. no. of occurrences; (a) BU/SU OBSB 3x3, (b) WW OBSB 3x3, (c) Antbirds 5x5

5. Discussion

Ideally, ornithologists would like a measure and methodology to compare calls or songs from different populations that allows them to say that the vocalizations are very similar or quite different. It may not be possible to exactly pinpoint where the threshold between similar and different lies. Such evidence is always taken in parallel with genetic and morphological data about the bird population to build a complete picture. We aim to identify regions of confidence for similarity and difference with a boundary region in between. This requires more development of the current approach. These initial results are promising however. The measure for the overlapped BNO populations is smaller than it is for the Antbird populations which are known to be diverse and from separate species. To give a reference point for this small measure, two non-overlapped portions of the DAB data were used to build histograms and the distance measure calculated for the two populations. This allowed a reference comparison of two populations with highly similar songs. The measure was 0.0518 for the 3X3 codebook and 0.3095 for the 5X5. Thus results of this order can be considered highly similar.

Moving to the 5X5 codebook made the distance between the AF and BS species the largest, suggesting greater diversity in their calls. However, the data in this set was the lowest quality data used and it is likely the BS data contained multiple subspecies as we could not get enough data from a single geographical location. Control over and access to data is an ongoing challenge in this development. Intuitively the use of the 5X5 codebook allows capture of a more complex micro-structure of pitch contour. Whether this is necessary for all bird population comparisons is not clear. For instance, the BNOs call is the simplest of the calls used in this paper hence the measure for the overlapped BNO populations is 0.3095. However, the data in this set was the lowest quality data used and it is likely the BS data contained multiple subspecies as we could not get enough data from a single geographical location.

A better representation of the pitch contour space using vector quantization is possible. This could replace the uniformly spaced codebook with a more flexible representation that can be tuned for a given species comparison. The current thresholds are quite high and may not allow a sufficiently nuanced division of the pitch microstructure space. Ornithologists need to know how many recordings they need to take in the field to later analyze bird vocalizations as such data is not easy to collect. Experiments to determine how stable the measure is with varying number of recordings are planned. Further experiments will also examine whether 70% and 90% thresholds are suitable, and optimize the relationship between overall codebook arrangement and system.

6. Conclusion

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


