Generalised Features for Bird Vocalisation Retrieval in Acoustic Recordings

Xueyan Dong #1, Jie Xie*2, Michael Towsey*3, Jinglan Zhang *4, Paul Roe*5

# Queensland University of Technology (QUT)
Queensland, Australia

1 xueyan.dong@hdr.qut.edu.au
2 j3.xie@hdr.qut.edu.au
3 m.towsey; jinglan.zhang; p.roe@qut.edu.au

Abstract— Bioacoustic monitoring has become a significant research topic for species diversity conservation. Due to the development of sensing techniques, acoustic sensors are widely deployed in the field to record animal sounds over a large spatial and temporal scale. With large volumes of collected audio data, it is essential to develop semi-automatic or automatic techniques to analyse the data. This can help ecologists make decisions on how to protect and promote the species diversity. This paper presents generic features to characterize a range of bird species for bird vocalisation retrieval. In the implementation, audio recordings are first converted to spectrograms using short-time Fourier transform, then a ridge detection method is applied to the spectrogram for detecting points of interest. Based on the detected points, a new region representation are explored for describing various bird vocalisations and a local descriptor including temporal entropy, frequency bin entropy and histogram of counts of four ridge directions is calculated for each sub-region. To speed up the retrieval process, indexing is carried out and the retrieved results are ranked according to similarity scores. The experimental results show that our proposed feature set can achieve 0.71 in term of retrieval success rate which outperforms spectral ridge features alone (0.55) and Mel frequency cepstral coefficients (0.36).

I. INTRODUCTION

Species diversity is critical to human beings because they can provide a variety of services, such as food, living material, and recreation. According to a recent Australian Biodiversity Strategy, more than 1700 species in Australia are known to be threatened and at risk of extinction due to the effects of human activities and environmental changes [1]. The conservation of species diversity becomes an urgent need.

A traditional way of species conservation is to conduct a survey based on field observation. This approach can achieve comprehensive results by conforming to standard protocols. However, these surveys are often carried out in a small geographic and temporal scale. An alternative mean is acoustic monitoring using sensors, which has been widely used in many studies [2-4]. Typically, sensors are placed in a wild area to record animal sounds. This method has great advantages in collecting data over large spatiotemporal scale. The collected sounds can be analysed multiple times and assist ecologists in understanding vocal species.

Acoustic sensors collect large volumes of audio data, which requires automated tools for processing. Current algorithms for automated animal sound analysis focus on species recognition and retrieval. The selection of discriminating features is the key to the success of these two tasks. The feature-based approach has two benefits: (1) the large amount of audio data can be reduced to a compact feature space; and (2) the selected features are useful for distinguishing among various species. These features can be adaptive for multiple class recognition or retrieval.

In previous studies of bird song recognition, features are often designed for describing a limited number of species or a particular type of bird sounds. However, in order to identify multiple bird species, a generic feature set is required. In addition, most of studies in the context focus on species classification [5-8] while few efforts have been put to birdsong retrieval in continuous acoustic sensor recordings.

In this paper, we present a generalised feature representation to characterize a wide range of bird species for vocalisation retrieval. The developed features are applied to a query-by-example retrieval system over a database of birdsong recordings collected in the field. This study makes three contributions: (1) a new way to detect a range of bird calls from environmental recordings, which is good for bird vocalisations; (2) a novel region representation for characterizing bird vocalisations of multiple species, which shows great benefits in differentiating short calls from complex calls; and (3) the application of developed features to retrieve bird vocalisations over continuous acoustic recordings, which is useful for detecting bird species’ presence or absence.

In addition, averaged Mel-frequency cepstral coefficients (MFCCs) and the derived statistics (ΔMFCC and ΔΔMFCC), which are widely used features in audio recognition, are computed for comparison. To the best of our knowledge, it is the first time that a generic feature set is explored to retrieve a range of bird vocalisations in continuous real-world recordings.

The reminder of the paper is organized as follow. Section II reviews related work. Section III discusses a signal detection method and a feature representation approach for bird vocalisation retrieval. Section IV reports the experimental results using our proposed features and baseline features. Conclusion is given in Section V.
II. RELATED WORK

For automatic birdsong analysis, signal detection (or segmentation) is a necessary step which aims to separate signals of interest from background noise. Many approaches attempt to achieve the goal based on time-frequency representation (spectrogram). A simple way is to set up an intensity threshold to select the sound of interest. Brandes’s work [5] shows that most animal calls are frequency modulated, which means different species, such as frogs, crickets, and birds, make calls in distinct frequency bands, therefore, applying an adaptive threshold for each frequency band is required. However, Neal et al. [9] point out that this threshold method is ineffective in segmenting field recordings where multiple sound sources are recorded. Thus they explore a binary classification method to differentiate between bird and non-bird events. However, this method requires an amount of training data, which is not useful when large training data is not available.

There are various features explored for representing bird sounds in automated species recognition. Spectral features (call bandwidth and spectral flatness) are extracted from spectrograms which are derived from the short-time Fourier transform (STFT) [7]. MFCC models offer a compact parametric representation of bird calls with broadband characteristics and harmonics [10, 11]. However, as pointed out by Somervuo et al. [12], cepstral coefficients misrepresent important pitch information which is equivalent to magnitudes of amplitude in the spectrogram and their suitability for many birdcalls is questionable. Since many bird calls consist of tonal structures, time-varying sinusoids are modelled from such a type of bird call. [13, 14]. Jančovič and Kókúr report that sinusoidal models provide a better representation of bird calls in field recordings than standard MFCCs widely used in speech recognition [15].

Spectrograms can be viewed as images (despite neither of the dimensions being spatial) and a range of image processing techniques have been applied to the problem of birdcall recognition. Two more recent examples are the MPEG angular radial transform [16] and Histograms of Oriented Gradients (HOG) [17]. Note that translation invariance (in frequency) and rotational invariance are not appropriate for characterizing spectral representations of bird calls and therefore the relevance of some image processing techniques must be questioned. HOG are successfully applied to acoustic signals by [19] for the determination of speaker gender in speech. HOG features are combined with other acoustic features in the bird call classification task of [17] but the contribution of the HOG features to the final result is not reported.

The feature extraction approaches presented in the reviewed studies are often designed for particular applications. Therefore, they are only appropriate for characterizing particular type of species, which are useful for species classification. In contrast, a retrieval system requires a more general method to allow arbitrary queries which may cover a wide range of bird species.

III. METHOD

A. Datasets

The dataset in the study is designed to validate the effectiveness of the proposed features for a birdcall retrieval system. The QUT eco-acoustic research group has collected over 24 terabytes of recordings of animal sounds from different fields over multiple years using acoustic sensors. In particular, this study focuses on the dataset collected from the Samford Ecological Research Facility (SERF), an open bush land located in 20 kilometers north-west of Brisbane CBD, Queensland, Australia. It contains five days (24 hours, 13th to 17th of Oct in 2010) × four sites recordings and corresponding annotation data. Wimmer et al. [4] report the details about how the recordings were collected. We use a subset of this dataset for the experiment.

The queries in this study include a representative range of 24 bird species and their names are listed in table I. Each species has five typical vocalizations as queries, so in total there are 120 queries in the query set. The selected bird vocalisations show distinctive structures and cover a range of call structures defined in the reviewed work [20], example

<table>
<thead>
<tr>
<th>No.</th>
<th>Species Name</th>
<th>Common Name</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Macropygia amboinensis</td>
<td>Brown Cuckoo-dove</td>
<td>BCD</td>
</tr>
<tr>
<td>2</td>
<td>Cacomantis variolosus</td>
<td>Brush Cuckoo</td>
<td>BCK</td>
</tr>
<tr>
<td>3</td>
<td>Lichmera indistincta</td>
<td>Brown Honeyeater</td>
<td>BHE</td>
</tr>
<tr>
<td>4</td>
<td>Burhinus grallarius</td>
<td>Bush Stone-curlew</td>
<td>BSC</td>
</tr>
<tr>
<td>5</td>
<td>Psophodes olivaceus</td>
<td>Eastern Whipbird</td>
<td>EWB</td>
</tr>
<tr>
<td>6</td>
<td>Eopsaltria australis</td>
<td>Eastern Yellow Robin</td>
<td>EYR</td>
</tr>
<tr>
<td>7</td>
<td>Rhipidura albiscapa</td>
<td>Grey Fantail</td>
<td>GFT</td>
</tr>
<tr>
<td>8</td>
<td>Colluricincla harmonica</td>
<td>Grey Shrike-thrush</td>
<td>GST</td>
</tr>
<tr>
<td>9</td>
<td>Pachycephala pectoralis</td>
<td>Golden Whistler</td>
<td>GWS</td>
</tr>
<tr>
<td>10</td>
<td>Philemon citreogularis</td>
<td>Little Friarbird</td>
<td>LFB</td>
</tr>
<tr>
<td>11</td>
<td>Myiagra rubecula</td>
<td>Leaden Flycatcher</td>
<td>LFC</td>
</tr>
<tr>
<td>12</td>
<td>Meliphaga lewinii</td>
<td>Lewins Honeyeater</td>
<td>LHE</td>
</tr>
<tr>
<td>13</td>
<td>Oriolus sagittatus</td>
<td>Olive-backed Oriole</td>
<td>OBO</td>
</tr>
<tr>
<td>14</td>
<td>Pachycephala rufiventris</td>
<td>Rufous Whistler</td>
<td>RFW</td>
</tr>
<tr>
<td>15</td>
<td>Trichoctilus haematodus</td>
<td>Rainbow Lorikeet</td>
<td>RLK</td>
</tr>
<tr>
<td>16</td>
<td>Chrysococcus lucidus</td>
<td>Shining Bronze-cuckoo</td>
<td>SBC</td>
</tr>
<tr>
<td>17</td>
<td>Cacatua galerita</td>
<td>Sulphur-crested Cockatoo</td>
<td>SCC</td>
</tr>
<tr>
<td>18</td>
<td>Zosterops lateralis</td>
<td>Silvereye</td>
<td>SVE</td>
</tr>
<tr>
<td>19</td>
<td>Myzornela sanguinolenta</td>
<td>Scarlet Honeyeater (call)</td>
<td>SHE1</td>
</tr>
<tr>
<td>20</td>
<td>Myzornela sanguinolenta</td>
<td>Scarlet Honeyeater (song)</td>
<td>SHE2</td>
</tr>
<tr>
<td>21</td>
<td>Pardalotus striatus</td>
<td>Striated Pardalote</td>
<td>SPD</td>
</tr>
<tr>
<td>22</td>
<td>Corvus orru</td>
<td>Torresian Crow</td>
<td>TRC</td>
</tr>
<tr>
<td>23</td>
<td>Melithreptus albogularis</td>
<td>White-throated Honeyeater</td>
<td>WTH</td>
</tr>
<tr>
<td>24</td>
<td>Lichenostomus chrysops</td>
<td>Yellow-faced Honeyeater</td>
<td>YFH</td>
</tr>
</tbody>
</table>
Fig. 1. Example spectrograms for each of 6 bird call classes that are (a) Eastern Yellow Robin (EYR)(two clicks), (b) Rufous Whistler (RFW)(a series of chirping), (c) Bush Stone-curlew (BSC)(a whistle), and (d) Sulphur crested-cockatoo (SCC)(shrieks), (e) Eastern whipbird (EBW)(whistle and click), and (f) Torresian Crow (TRC)(stack harmonics).

spectrograms of call classes are displayed in Fig. 1. In order to cover representatives of the selected species from the 20-days’ recordings, the query set is chosen from different sites and different time. A query here is prepared by manually specifying a region which contains a bird vocalisation in the spectrogram.

For the search database, we ensure that recordings in the query set are excluded. In the end, we chose one day of recordings, on the 13th Oct in north-east of the recording site. 80 species are present in the recordings. The recordings are cut into one-minute segment for simple analysis. Each segment is formatted with a sampling rate of 22,050 Hz and 16-bit resolution.

B. The flowchart of the retrieval system

There are five major procedures in the designed retrieval system, which is shown in Fig. 2. First, all audio files are converted to spectrograms using STFT. Then spectral ridge detection is applied to the spectrograms. The detected ridges are used to parametrize into feature vectors. Since our system aims to process a large amount of audio files, indexing is added to improve the retrieval speed. In the end, the system retrieves similar bird vocalisations to the query. The detail of each procedure is discussed in the following subsections.

1) Spectrogram Preparation: Spectrograms are generated using STFT with a Hamming window of 512 samples (23ms) and 50% window overlap. We denote spectral values by $X(t,f)$, where $t$ represents a time frame and $f$ indexes a discrete frequency bin. These spectral pairs correspond to pixels of spectrogram image. Spectral amplitude values are converted to decibels (dB) using $\text{dB} = 20\log_{10}(X)$. To reduce background noise, we apply a noise removal algorithm developed by Towsey et al. [21] which calculates a separate decibel threshold for each frequency bin assuming an additive noise model.

2) Spectral Ridge Detection: Most bird calls exhibit spectral ridge components that can be seen in Fig. 2. The structure tends to show ridge characteristics in terms of the intensity values on the spectrogram. Therefore we first attempts the ridge detection method designed by Dong et al. [22] to identify portions of the spectrogram that show ridge characteristics. In this method, ridge pixels are detected by convolving each prepared spectrogram with four masks, one mask for each ridge direction. Here we employ the set of masks for the directions $0, \frac{\pi}{4}, \frac{\pi}{2},$ and $3\frac{\pi}{4}$ radians. A pixel in the spectrogram is assigned a ridge direction corresponding to the mask yielding maximum convolution score only if the score exceed a threshold of 6.0 dB.

Through the experiment, we found Dong’s ridge detection is not suitable for detecting calls that show shrieks (a block shape). This case is quite common in bird songs collected in the wild environment because of echo effect or birdsongs themselves. Fig. 1 (a), (b) are bird calls exhibiting shadow due to environmental effects. Therefore we modify the ridge detection method for our application.

To address the problems with shriek calls, scale factor ($\sigma$) is considered to modify the previous ridge detection. $\sigma$ is designed to compress the spectrogram along time and frequency directions so that ridges can stand out for the shriek calls. We test the values of $\sigma$, 0.0, 0.125, 0.25, 0.5, and determine 0.25 is the optimum to implement the spectrogram compression to derive ridges. An example of spectrogram compression can be seen in Fig. 3 (c) and (d). The derived ridges then add to the original ridges obtained in the uncompressed spectrogram, see Fig. 3 (e).

3) Feature Extraction: A query is a section of audio with arbitrary duration and frequency bounds, represented as a rectangular section of spectrogram. Examples are shown in Fig. 1.
For a Scarlet Honeyeater call (a rectangle with grey shading), it is divided into 18 regions, each of which is shown as a square in the figure. Each square has an index which refers to a 6-dim feature set derived from ridges inside the square. The squares surrounding the call belong to a buffer zone.

Ridges are found in the buffer zone of a candidate instance, the similarity score decreases due to mismatching regions happened in the buffer zone.

To describe each region, we calculate a six dimensional feature vector: 1. temporal entropy (1-D); 2. frequency bin entropy (1-D); 3. a histogram of four ridge directions (4-D);

1. **Temporal entropy** ($H_T$): The ridge magnitudes are summed frame-wise over all frames in the region and the N values are normalized to unit sum. $HT$ is calculated as:

\[
H_T = - \sum p_i \log_2 p_i \quad \text{where} \quad i \in [1, N]
\]

2. **Frequency bin entropy** ($H_f$): Similar to the calculation of $H_r$, except that the ridge magnitudes are summed bin-wise over all bins in the region. $H_f$ and $H_r$ can describe the spatial distribution of spectral ridges in a region.

3. **Histogram of counts of four ridge directions** ($HoRC4$): To further describe the local property of each region, we calculated a histogram of four ridge directions that is inspired by Histogram of Oriented Gradient [18]. Here a four-dimensional vector is derived from the counts of region cells belonging to ridges having direction 0, $\pi/4$, $\pi/2$ and $3\pi/4$ rather than the magnitude used by Dalal and Triggs [18]. The histogram values are normalized to $[0,1]$. Whereas the entropy features describe the spatial distribution of ridge cells within a region, this feature describes the distribution of ridge directions.

4. **Indexing**: Indexing here is to pre-calculate features for speeding up the matching process on a large audio collection. An audio file ($Di$) in the search database is represented by a set of regions ($r$). The matrix of spectrogram for $Di$ is divided into overlapping regions along time frames. This operation allows the variations of bird calls in frames. Here the shift for neighboring regions is chosen as half of the region size, 5 frames. So $Di = \{r_1, r_2, r_3, \ldots, rm\}$ ($m$ = the total number of regions in $Di$). In this step, a region can be parametrized as $n = \{f, t, fc\}$ where $f$ refers to the low frequency bin index, $t$ is start frame index, and $fc$ is the 6-d features. The generated index for each $Di$ is stored into a csv file and the index item is distributed as a matrix.

5. **Retrieval**: When searching for potential candidates, the query grid is applied to the generated indexing. The search results in approximately 100 matching regions from each one-minute recording in the database. Such an amount of candidates are determined by a filtering step that aims to eliminate the regions in which 50% of the candidate sub-regions underlying the query grid do not contain ridges.

Similarity matching is achieved by using K-NN ($K = 1$) which means we only count the individual highest neighbour when determining retrieved calls. Here a similarity score for each candidate is obtained by calculating the overall similarity ($S$) between a query call and a candidate region. $S$ is derived using weighted average score calculated from corresponding regions within the regions. Since empty regions lead to bias to the score, we give less weight (0.2) for them but more
weight (0.8) to ridge regions. Consider the size of query would affect the score, therefore, the final score is computed through dividing by maximum score for the exact match.

The retrieved candidate instances are ranked by similarity score. The highest similarity score is 1.0, which means exact match. In fact, exact match seldom happen due to complexity of birdcalls in field recordings.

6) Baseline method: In the experiment, MFCCs, ΔMFCC, and ΔΔMFCC are used as the baseline for feature comparison. The method for MFCCs extraction is a modified version of an algorithm developed by Lee et al. [24]. According to the time domain boundary of a bird call region, MFCCs are first extracted from each frame of the acoustic event. Then, the averaged MFCCs of all frames within the bird call are calculated as (2).

\[ f_m = \frac{\sum_{k} c_k}{K}, \quad m \in [0, L - 1] \quad (2) \]

where \( f_m \) is the \(m\)th MFCCs, \( K \) is the number of frames for one event, \( c_m \) is DCT result of each filtered amplitude spectrum. \( L \) is the number of feature vector for each frame, and here it is 13. The final feature is represented by the normalised MFCCs, which is shown in (3).

\[ MFCCs = \frac{f_m - f_{mmin}}{f_{mmax} - f_{mmin}} \quad (3) \]

To further explore the performance of our spectral ridge features (termed as ISR), ΔMFCCs and ΔΔMFCCs (parameters commonly used in automatic speech recognition) as well as the spectral ridge features (SR) derived from original ridge detection and region representation without a buffer zone are also computed for comparison.

7) Performance evaluation: To evaluate the retrieval performance, the top N success rate is calculated, which reflects the percentage of all queries achieves correct matches among top N. Notice GFT and SBC are not present in the recording, so the total query count calculated here is 110 rather than 120.

IV. RESULTS

Both spectral ridge feature sets outperform the MFCCs feature sets. In addition, the SR (CB) feature set (yielding a final accuracy of 71%) outperforms the standard SR feature set (Table II). Note that this accuracy figure is derived from the presence of at least one correct response in the top five returns. The addition of Delta features to the standard MFCC features improves performance a little and so the remaining MFCC benchmark results are shown only for MFCCs plus Delta features.

**TABLE II**

<table>
<thead>
<tr>
<th>Success Rate(N)</th>
<th>MFCCs</th>
<th>MFCC + Δ</th>
<th>MFCC + ΔΔ</th>
<th>SR</th>
<th>SR(CB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1</td>
<td>0.20</td>
<td>0.20</td>
<td>0.19</td>
<td>0.27</td>
<td>0.39</td>
</tr>
<tr>
<td>Top 3</td>
<td>0.32</td>
<td>0.31</td>
<td>0.22</td>
<td>0.43</td>
<td>0.56</td>
</tr>
<tr>
<td>Top 5</td>
<td>0.33</td>
<td>0.36</td>
<td>0.35</td>
<td>0.55</td>
<td>0.71</td>
</tr>
</tbody>
</table>

It is instructive to compare performance on individual species (Table III). An accuracy of 1.0 indicates that the correct response is present in the top five returns for each of the five queries. Note that no GBT or SBC calls are present in the 24 hour recording and consequently the correct responses to the ten queries for these two species are not to return any matches. These two species are shown at the bottom of Table III and it will be observed that no matches are found with a similarity score less than a threshold of 0.5. A threshold on returns is useful if the object of the queries is to determine species presence/absence.

It is significant that the MFCCs feature set performs well on those birds having broad-band calls (e.g. RLK, TRC and SHE1) and poorly on the remainder. This weakness can be explained by the investigation that the cepstral transform loses frequency information which is particularly important for distinguishing bird species. In summary, MFCCs features are not suitable for representing bird calls.

Spectral ridge features are improved when calculated after ridge compression (SR(C)) and in conjunction with a buffer zone around the query call (SR (CB)). The best retrieval result is 71%. The ridge compression technique alone (SR) obtains poor performance in detecting parrot shriek structures, such as SCC and RLK but this problem is addressed by the spectrogram compression process. Another drawback in SR is dealing with the queries of short calls, like BCD, OBO and BSC, because they can match to similar components of larger
calls. This problem can be solved by surrounding the query with a “buffer zone” in which no acoustic energy should appear. This idea improves retrieval on four of the 24 species. In summary, the extraction of spectral ridge features (SR) combined with ridge compression (SR (C) to detect diffuse calls) and the use of a surrounding combining with compression (SR (CB)) achieves the highest retrieval rate (Table III).

Nevertheless the method still performs poorly for three species, LFC, BHE and OBO. The calls of LFC and BHE are highly variable which makes accurate retrieval difficult. OBO is a short, narrow bandwidth call and its queries not only matched parts of other calls but also background noise. It should be recalled that the 24-hour recording used in this study contained the calls of 80 bird species and no editing was done to remove the many forms of background noise that occur in the environmental recordings.

V. CONCLUSION

This paper presents robust spectral ridge features for bird call retrieval over continuous recordings collected from the natural environment. The proposed features work well in characterizing a wide range of bird species. The experimental results demonstrate that they perform better than previous spectral ridge features and MFCCs features alone. A strong test of our retrieval system is that it performs well on a 24-hour recording containing undefined environmental noise as well as some 60 bird species that had not been seen prior to development of the system. This is particularly significant since a bird call classification system must be retrained for each new occurrence of a species and even to accommodate regional variation. The flexibility of our system makes it particularly useful for real-world ecological studies as opposed to systems that work with fixed databases. For future work, we wish to address the problem of simultaneous calls from different species having overlapping bandwidths and conduct more comparative studies with other existing methods.

REFERENCES