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*
Queensland, Australia

[1] xueyan.dong@hdr.qut.edu.au

[2] j3.xie@hdr.qut.edu.au

[3]m.towsey; [4]jinglan.zhang; [5]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 vocalisation retrieval. In the implementation, audio recordings are first converted to spectrograms using short-time Fourier transform, then a modified 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 experiment results show that our proposed feature set can achieve 0.71 in term of retrieval success rate which outperforms spectral ridge (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 especially good for bird vocalisations having block shape structures; (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 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 birdcalls 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 Kökuer reported 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

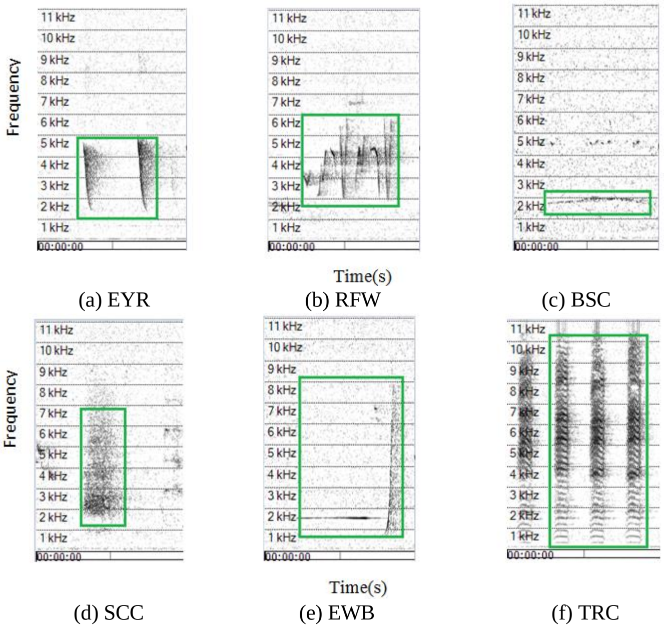
The dataset in the study is designed to validate the effectiveness of the proposed features for a birdcall retrieval system. The QUT eco-acoustic research group has collected over 24 terabytes of recordings of animal sounds from different fields over multiple years using acoustic sensors. In particular, 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 to17th 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.

Table I

Bird species in the study

No.	Species Name	Common Name	Code
1	*Macropygia amboinensis*	Brown Cuckoo-dove	BCD
2	*Cacomantis variolosus*	Brush Cuckoo	BCK
3	*Lichmera indistincta*	Brown Honeyeater	BHE
4	*Burhinus grallarius*	Bush Stone-curlew	BSC
5	*Psophodes olivaceus*	Eastern Whipbird	EWB
6	*Eopsaltria australis*	Eastern Yellow Robin	EYR
7	*Rhipidura albiscapa*	Grey Fantail	GFT
8	*Colluricincla harmonica*	Grey Shrike-thrush	GST
9	*Pachycephala pectorails*	Golden Whistler	GWS
10	*Philemon citreogularis*	Little Friarbird	LFB
11	*Myiagra rubecula*	Leaden Flycatcher	LFC
12	*Meliphaga lewinii*	Lewins Honeyeater	LHE
13	*Oriolus sagittatus*	Olive-backed Oriole	OBO
14	*Pachycephala rufiventris*	Rufous Whistler	RFW
15	*Trichoglossus haematodus*	Rainbow Lorikeet	RLK
16	*Chrysococcyx Iucidus*	Shining Bronze-cuckoo	SBC
17	*Cacatua galerita*	Sulphur-crested Cockatoo	SCC
18	*Zosterops laterails*	Silvereye	SVE
19	*Myzonmela sanguinolenta*	Scarlet Honeyeater (call)	SHE1
20	*Myzonmela sanguinolenta*	Scarlet Honeyeater (song)	SHE2
21	*Pardalotus striatus*	Striated Pardalote	SPD
22	*Corvus orru*	Torresian Crow	TRC
23	*Melithreptus albogularis*	White-throated Honeyeater	WTH
24	*Lichenostomus chrysops*	Yellow-faced Honeyeater	YFH

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 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.



(a) EYR (b) RFW (c) BSC

(d) SCC (e) EWB (f) TRC

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).

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 dB = 20log10(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.

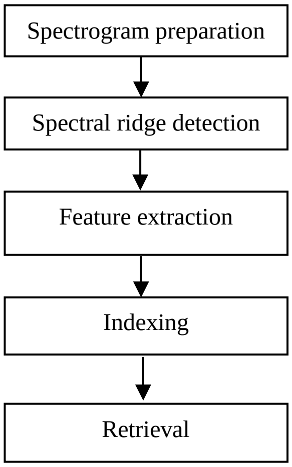


Fig. 2. Flowchart of the retrieval system.

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, π/4, π/2, and 3π/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 (σ) is considered to modify the previous ridge detection. σ 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 σ, 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.

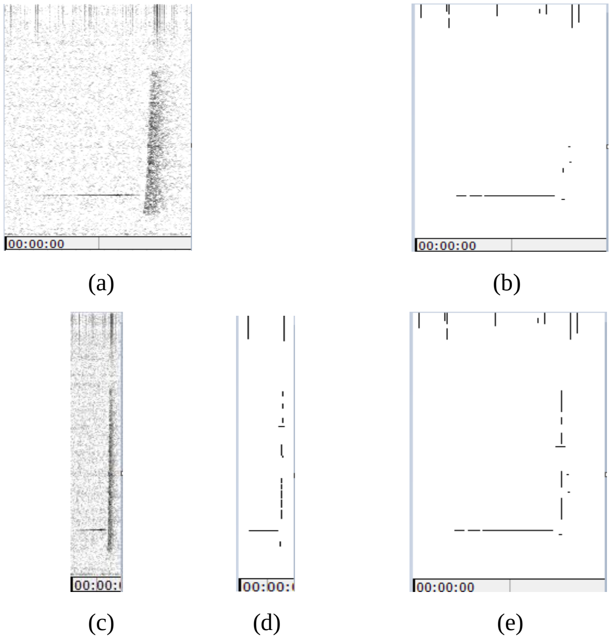


Fig.3. Ridge detection results on different scales of spectrograms (a) Noise-reduced spectrogram of an Eastern Whipbird call containing a shadow; (b) ridge detection on original spectrogram (c) time compressed spectrogram. (d) ridge detection on time-compressed spectrogram. (e) restored-scale spectrogram. The duration of the call is 2 seconds. The call ranges from 1700~7500 Hz. The vertical ridges appearing above 8000 Hz are detected due to MP3 artefacts.

To capture local variations in bird calls, we develop a normalized block descriptor. A bird call is divided into a grid of non-overlapping square blocks of size 11×11, termed as regions. The size of bird call can be arbitrary and the number of 11x11 regions generated depends on the size. Finally the call is characterized as the vector of all region features within the call.

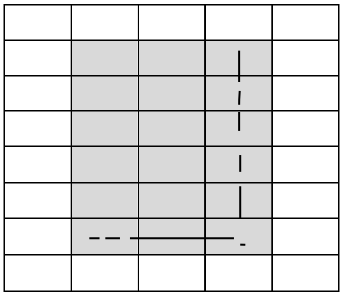


Fig.4. Region representation of spectral ridge features for a simplified Scarlet Honeyeater call.

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.

As pointed out by Arganat [23], short calls can cause an issue during the identification of multiple species because they lack of distinct properties compared to complex calls. The query set in the study contains many short calls which might be confused with large calls. To avoid this problem, we add a buffer zone to the actual bird call, which is shown in Fig. 4. In the buffer zone, each region should contain no ridges. When ridges are found in the buffer zone of candidate instance, the similarity score decreases due to mismatching regions happened in buffer zones.

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:

$$H_t = -\sum p_i log_2 p_i \text{ Where } i \in [1, N] \qquad (1)$$

2. Frequency bin entropy (H_f): Similar to the calculation of H_t except that the ridge magnitudes are summed bin-wise over all bins in the region. H_t and H_f 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, π/4, π/2 and 3π/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 (D_i) in the search database is represented by a set of regions (r). The matrix of spectrogram for D_i is divided into overlapping regions along time frames. This operation allows the variations of bird calls in frames. Here the shift for neighbouring regions is chosen as half of the region size, 5 frames. So $D_i = \{r1, r2, r3, \dots\ rm\}$ (m = the total number of regions in D_i). 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 f_c is the 6-d features. The generated index for each D_i 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.

IV. EXPERIMENT

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).

$$f_m = \frac{\sum_{i=1}^{K} C_m^i}{K}, \;\; m \in [0, L-1] \qquad (2)$$

where f_m is the m^{th} MFCCs, K is the number of frames for one event, C_m^i 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).

$$MFCCs = \frac{f_m - f_m^{min}}{f_m^{max} - f_m^{min}} \qquad (3)$$

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.

To evaluate the retrieval performance, the success rate is calculated, which reflects how many queries obtain correct retrieval within top rank. Notice *GFT* and *SBC* are not present in the recording, so the total query count calculated here is 110 rather than 120. From Table II, we find that spectral ridge methods, both SR and ISR, perform better than MFCCs based features. Among these feature sets, our improved spectral ridge (ISR) method achieves best result which yields correct retrievals for 71% of all queries within top five. The spectral ridge method obtains 55%. In contrast, MFCCs features obtain lower rate (around 35%), this illustrates they are not suitable for detecting birdcalls in field recordings. The three MFCCs features perform better for detecting broadband calls, e.g. *RLK*, *TRC* and *SCC*. But they show poor performance in other birdcalls. One reason is that they capture information in the whole frequency band within an interval of bird call such that it is sensitive to noise and insufficient to find similar calls when overlapping calls happen in time. In addition, we find that there is little difference among MFCCs, ΔMFCCs and ΔΔMFCCs. This reports that MFCCs is not suitable for representing bird calls.

To examine the performance for detecting species, the average accuracy within top five for four feature sets is computed and shown in table III. Since we have five queries for each species, the accuracy value can be 0.0, 0.2, 0.4, 0.6, 0.8 or 1.0. When accuracy is 1.0, it means that all queries obtain correct retrievals.

TABLE II

SUCCESS RATE FOR VARIOUS FEATURES

Success Rate(N)	**MFCCs**	**ΔMFCCs**	**ΔΔMFCCs**	**SR**	**ISR**
Top 1	0.20	0.20	0.19	0.27	0.39
Top 3	0.32	0.31	0.22	0.43	0.56
Top 5	0.33	0.36	0.35	0.55	0.71

To determine the species presence or absence, a threshold (t) is set for similarity score (s). If s is lower than t, it indicates that the querying species is absent. When t = 0.5, The retrieval results demonstrate that the improved spectral ridge (SR+C, and SR+C+B) and spectral ridge (SR) can detect the majority of species (21) presence except for *LFC*, *GFT*, and *SBC*. *LFC* is actually record in the search database but the calls have many variations, which cause retrieval errors given the queries. In contrast, Both MFCCs and ΔMFCCs identify 14 bird species but ΔΔMFCCs find 15 species. These results show that ISR achieves best performance in detecting species in the database.

TABLE III

AVERAGE ACCURACY AT TOP FIVE FOR 24 BIRD SPECIES IN THE STUDY (- INDICATING SPECIES NOT EXISTING IN THE DATABASE)

Species	**Accuracy (*C* = 5)**			
	ΔMFCCs	**SR**	**SR+C**	**SR+C+B**
BCD	0.0	0.4	0.4	**0.8**
BCK	0.0	0.2	0.2	**0.6**
BHE	0.2	0.2	0.2	0.2
BSC	0.0	0.2	0.2	**0.8**
EWB	0.2	1.0	1.0	1.0
EYR	0.0	0.8	0.8	0.8
GFT	-	-	-	-
GST	0.0	0.8	0.8	0.8
GWS	0.0	0.6	**1.0**	1.0
LFB	0.4	0.2	0.2	**0.6**
LFC	0.0	0.0	0.0	0.0
LHE	1.0	1.0	1.0	1.0
OBO	0.2	0.2	0.2	0.2
RFW	0.6	0.8	0.8	0.8
RLK	1.0	0.2	**0.6**	0.6
SBC	-	-	-	-
SCC	0.6	0.2	**1.0**	1.0
SVE	0.2	0.4	0.4	0.4
SHE1	0.8	1.0	1.0	1.0
SHE2	0.6	1.0	1.0	1.0
SPD	0.6	1.0	1.0	1.0
TRC	1.0	1.0	1.0	1.0
WTH	0.0	0.2	**0.4**	0.4
YFH	0.4	0.6	0.6	0.6
Average	**0.36**	**0.55**	**0.63**	**0.71**

SR obtains poor performance in detecting the shriek structures, such as *SCC* and *WTH*. Compression process can address the problem as the average accuracy for these species is higher than the ones obtained by SR, see the bold values in the column of SR + C (refers to compression). Another drawback in SR is dealing with short calls, like *BCD* and *BSC*, as they are easily confused with large patterns of bird calls. *BCD* tends to be confused with noise and *TRC*. *BSC* particularly confuses with *EWB* as both of them contain whistle (a horizontal line in the spectrogram). The spectral ridge method (SR) combining with compression (C) and buffer (B) zone can address the situation, which is reflected in the last column of Table III.

The most difficulty for our modified method applying to birdcall retrieval in field recording is the variations in bird calls, like *OBO* and *LFC*. They are found either in birdcalls themselves or background sounds. This demonstrates that different individual species may produce different calls or their calls may be captured differently by acoustic sensors. Another issue is the confusion with untargeted species.

V. Conclusion

This paper presents spectral ridge features for bird call retrieval over continuous recordings collected in natural environment. The proposed features work well in characterizing a wide range of bird species. The experimental results demonstrate that they perform better than SR and MFCCs features. As discussed in the experiment section, MFCCs are not appropriate for describing most birdcalls because they are sensitive to the background sounds. In terms of dealing with field recordings, the developed feature representation has great suitability to differentiate short calls with large patterns of bird songs using a buffer zone. Another advantage is that our proposed features can detect calls that are made simultaneously. The presented birdcall retrieval system can assist ecologists in discovering the presence or absence of species at a particular site.

Environmental acoustic data is difficult to analyse due to their complexity and varieties in bird species. One limitation of the developed features is that they are not sufficient for dealing with calls overlapping in frequency. In the future, approaches that address the limitation will be explored.

References

[1] A. H. Arthington and J. Nevill, "Australia's Biodiversity Conservation Strategy 2010–2020: scientists' letter of concern," *Ecological Management & Restoration,* vol. 10, pp. 78-83, 2009.

[2] R. Bardeli, D. Wolff, and M. Clausen, "Bird song recognition in complex audio scenes," *Computational Bioacoustics for Assessing Biodiversity. Proc. of the Internat. Expert Meeting on IT-based Detection of Bioacoustical Patterns. BfN-Skripten,* pp. 93-101, 2008.

[3] K.-H. Frommolt, K.-H. Tauchert, and M. Koch, "Advantages and disadvantages of acoustic monitoring of birds–realistic scenarios for automated bioacoustic monitoring in a densely populated region," *Computational Bioacoustics for Assessing Biodiversity. Proc. of the Internat. Expert Meeting on IT-based Detection of Bioacoustical Patterns. BfN-Skripten*, vol. 234, pp. 83-92, 2008.

[4] J. Wimmer, M. Towsey, P. Roe, and I. Williamson, "Sampling environmental acoustic recordings to determine bird species richness," *Ecological Applications,* 2013.

[5] T. S. Brandes, "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,* vol. 16, pp. 1173-1180, 2008.

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

[7] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," *The Journal of the Acoustical Society of America,* vol. 103, p. 2185, 1998.

[8] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric Representations of Bird Sounds for Automatic Species Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, pp. 2252-2263, 2006.

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

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 28, pp. 357-366, 1980.

[11] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE,* vol. 81, pp. 1215-1247, 1993.

[12] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, pp. 2252-2263, 2006.

[13] A. Harma and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04)*, 2004, pp. V-701-4 vol. 5.

[14] P. Jančovič, M. Köküer, and M. Russell, "Bird species recognition from field recordings using HMM-based modelling of frequency tracks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 8252-8256.

[15] P. Jančovič and M. Köküer, "Automatic detection and recognition of tonal bird sounds in noisy environments," *EURASIP Journal on Advances in Signal Processing,* vol. 2011, p. 982936, 2011.

[16] C.-H. Lee, S.-B. Hsu, J.-L. Shih, and C.-H. Chou, "Continuous Birdsong Recognition Using Gaussian Mixture Modeling of Image Shape Features," *IEEE Transactions on Multimedia,* vol. 15, pp. 454-464, 2013.

[17] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *The Journal of the Acoustical Society of America,* vol. 131, pp. 4640-4650, 2012.

[18] A. M. Selvan and R. Rajesh, "Spectral histogram of oriented gradients (SHOGs) for Tamil language male/female speaker classification," *International Journal of Speech Technology,* vol. 15, pp. 259-264, 2012.

[19] Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International,* vol. 18, pp. S163-S173, 2008.

[20] M. Towsey, J. Wimmer, I. Williamson, and P. Roe, "The use of acoustic indices to determine avian species richness in audio-recordings of the environment," *Ecological Informatics,* vol. 21, pp. 110-119, 2014.

[21] X. Dong, M. Towsey, J. Zhang, J. Banks, and P. Roe, "A Novel Representation of Bioacoustic Events for Content-Based Search in Field Audio Data," in *2013 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2013, pp. 1-6.

[22] I. Agranat, "Automatically Identifying Animal Species from their Vocalizations," presented at the Fifth International Conference on Bio-Acoustics, Holywell Park, 2009.

[23] C.-H. Lee, C.-H. Chou, C.-C. Han, and R.-Z. Huang, "Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis," *Pattern Recognition Letters,* vol. 27, pp. 93-101, 2006.