MURAI Common Malaysian Birds Sound Dataset

Muhammad Mun'im Ahmad Zabidi[1]
School of Electrical Engineering
Universiti Teknologi Malaysia
Skudai, Johor, Malaysia
[1]munim@utm.my

Rohana binti Mahmud[2], Mohd. Yamani Idna bin Idris[3]
Faculty of Computer Science and Information Technology
University of Malaya
Kuala Lumpur, Malaysia
[2]rohanamahmud@um.edu.my, [3]yamani@um.edu.my

***Abstract*—The potential applications of automatic species detection and classification of birds from their sounds are many (e.g., ecological research, biodiversity monitoring, archival)**

***Keywords*—bird sound detection, urban sound detection**

I. Introduction

Current state-of-the-art bird sound classifiers can recognize up to 1000 species but use big convolutional neural networks (CNNs) which require considerable compute and storage resources. To perform wok on low-complexity bird clasifiers, a small dataset is sufficient. Furthermore, a dataset customized for common Malaysian birds would be beneficial for local fieldwork.

II. Related Work

Similar works on small datasets have been done for urban sound detection and keyword spotting tasks.

A. Urban8k

To address these issues we present a taxonomy of urban sounds and a new dataset, UrbanSound, containing 27 hours of audio with 18.5 hours of annotated sound event occurrences across 10 sound classes. The challenges presented by the new dataset are studied through a series of experiments using a baseline classification system [1]. Urban8k contains 8732 audio files of urban sounds, 27 hour hours of audio with 18.5 hours of annotated sound event occurrences across 10 sound classes.

Meta-data included:

- **slice_file_name**:
  The name of the audio file. The name takes the following format:
  - `[fsID]-[classID]-[occurrenceID]-[sliceID].wav`, where:
  - `[fsID]` = the Freesound ID of the recording from which this excerpt (slice) is taken
  - `[classID]` = a numeric identifier of the sound class (see description of classID below for further details)
  - `[occurrenceID]` = a numeric identifier to distinguish different occurrences of the sound within the original recording
  - `[sliceID]` = a numeric identifier to distinguish different slices taken from the same occurrence
- **fsID**:
  The Freesound ID of the recording from which this excerpt (slice) is taken
- **start**:
  The start time of the slice in the original Freesound recording
- **end**:
  The end time of slice in the original Freesound recording
- **salience**:
  A (subjective) salience rating of the sound. 1 = foreground, 2 = background.
- **fold**:
  The fold number (1-10) to which this file has been allocated.
- **classID**:
  A numeric identifier of the sound class:

  0 = air_conditioner
  1 = car_horn
  2 = children_playing
  3 = dog_bark
  4 = drilling
  5 = engine_idling
  6 = gun_shot
  7 = jackhammer
  8 = siren
  9 = street_music

- **class**:
  The class name: air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren, street_music.

B. Google Speech Commands Dataset

The Speech Commands dataset is a standard training and evaluation dataset for a class of simple speech recognition tasks. Its primary goal is to provide a way to build and test small models that detect when a single word is spoken, from a set of ten or fewer target words, with as few false positives as possible from background noise or unrelated speech. This task is often known as keyword spotting [2].

All utterances are set to one second, 16KHz WAV. Single words are spoken in isolation, rather than as part of a sen-

TABLE I
TWO COMPACT DATASETS.

Dataset	#clips	Classes	Size	Sampling rate	Clip length (s)
Urban8k	8,732	10	27 hours	Various	Various < 4 s
Speech Commands	105,829	35	3.8GB	16 KHz WAV	1 s

tence, since this more closely resembles the trigger word task targeted.

The final dataset consisted of 105,829 utterances of 35 words, broken into the categories and frequencies shown in Table 1. Each utterance is stored as a one-second (or less) WAVE format file, with the sample data encoded as linear 16-bit single-channel PCM values, at a 16 KHz rate. There are 2,618 speakers recorded, each with a unique eight-digit hexadecimal identifier. The uncompressed files take up approximately 3.8 GB on disk, and can be stored as a 2.7GB gzip-compressed tar archive.

A few papers described the use of this dataset. ARM devised the CMSIS-NN library new optimized implementation of neural network operations for ARM microcontrollers, and uses Speech Commands to train and evaluate the results [3]. Reference [4] demonstrates how combining the dataset and UrbanSounds [1] can improve the noise tolerance of recognition models. Reference [5] shows how approaches learned from ResNet can produce more efficient and accurate models. Reference [6] investigates alternatives to traditional feature extraction for speech and music models. Google uses a small CNN to test the dataset [7].

III. BIRD DATASETS

In this section, we list the latest and most frequently used bird sound databases.

A. NIPS4Bplus

NIPS4Bplus is a 135 MB zip file dataset. The accompanying file contain temporal annotations for the recordings that comprised the training set of the 2013 Neural Information Processing Scaled for Bioacoustics (NIPS4B) challenge for bird song classification. Created by Morfi et al. [8]. Link: https://doi.org/10.6084/m9.figshare.6798548

B. Birdvox-full-night

The BirdVox project uses nine acoustic sensors near Ithaca, NY, USA, for monitoring avian migration. Full BirdVox data is 7k hours. The dataset containing some temporal and frequency information about flight calls of nocturnally migrating birds [10]. BirdVox-full-night only focuses on avian flight calls, a specific type of bird calls, that usually have a very short duration in time. The temporal annotations provided for them do not include any onset, offset or information about the duration of the calls, they simply contain a single time marker at which the flight call is active. Additionally, there is no distinction between the different bird species, hence no specific species annotations are provided, but only the presence of flight calls through the duration of a recording is denoted [8].

C. Bird-DB

Bird-DB, a publicly accessible relational database system that contains audio files of bird songs and their annotations for the phrase types they comprise [11].

Contains over 1000 recordings featuring individuals from more than 30 different bird species pertaining to the California and Western Australia regions. We acquired metadata about each of these audio recordings, modeled after a simplified version of the Macaulay Library metadata records.

Birds-DB currently has 428 files that have been annotated and several hundred more that have not been annotated yet. It currently requires a few terabytes of storage to include all the recordings. Link http://taylor0.biology.ucla.edu/birdDBQuery/

D. CLO-43SD

Most flight calls have a duration of less than 150 ms, and training a model on a full 1 second clip, at least 850 ms of which might not contain relevant data, could result in most training data not containing the relevant signal (the flight call). To account for this, we trimmed automatically every clip to 150 ms, centering on the middle of the original file with the assumption that a flight call was most likely to be centered around the precise time at which the template detector was triggered. In a random inspection of a small subset of 100 clips from the dataset, we found that all inspected clips still contained the flight call after the trimming procedure [12]. One of several dataset from BirdVox, a collaboration between the Cornell Lab of Ornithology and NYU's Music and Audio Research Laboratory (https://wp.nyu.edu/birdvox/codedata/#datasets).

E. Macaulay Library

The Macaulay Library is the world's largest and oldest scientific archive of biodiversity audio and video recordings. The Library is part of Cornell Lab of Ornithology of the Cornell University.

F. Tierstimmernarchiv

Tierstimmernarchiv (The Animal Sound Archive) at the Museum für Naturkunde in Berlin contains ca. 120,000 of over 2,500 species of animals. Over 16,000 recordings are publicly available [16].

IV. BIRD AUDIO DETECTION 2018

The Bird Audio Detection (BAD) challenge 2018 [13] is Task 3 of the DCASE Challenge. It is expanded from BAD 2016/2017. The dataset consists of 10 second-long audio recordings collected from real-life environments by e.g.

TABLE II
LATEST AND MOST FREQUENTLY USED DATASETS [8], [9].

Dataset Name	Location	#recs	#classes	Species tags	Annotations	Duration	Labelling
NIPS4Bplus [8]	France	687	87	Yes	Yes	1 h	Multi
LifeClef(BirdClef) 2019	Xeno-canto	50,000	659	Yes	No	350 h	
LifeClef(BirdClef) 2018	Xeno-canto	48,843	1500	Yes	No	68 h	
BirdVox-full-night [10]	New York	6	25	No	Yes	60 h	points in time
Bird-DB [11]	California/WA	428	100	Yes	Yes	—	Full syntax
CLO-43SD [12]	New York	5,428	43	No	No	—	Single
BirdVox-DCASE-20k	New York	20,000	N/A	No	No	55 h	Single
freefield1010 [13]	Freesound	7,690	N/A	No	No	21 h	Single
Warblr [13]	UK	10,000	N/A	No	No	28 h	Single
Chernobyl [13]	Russia	6,620	N/A	No	No	18 h	Single
PolandNFC [14]	Poland	4,000	N/A	No	No	11 h	Single
BAD 2018 DCASE Task 3 [13]			N/A	No	No	68 h	
Xeno-canto	Worldwide	573,000	10,238	Yes	No	9720 h	Citizen
Macaulay Library [15]	Worldwide	175,000+	6,196	Yes	No		
Berlin Museum für Naturkunde [16]		16,000	2,500	—	—	—	All animals

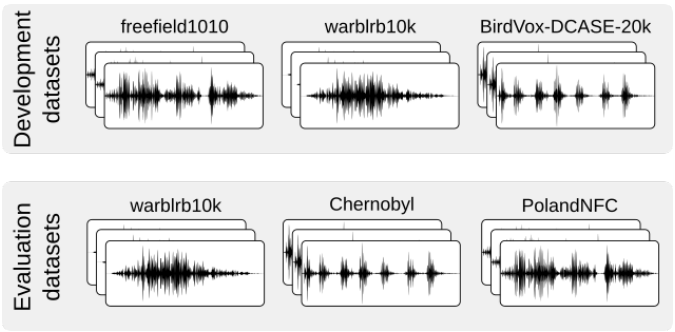


Fig. 1. DCASE 2018 Task 3 use of datasets.

crowd- sourcing from a mobile app for bird audio classification, and individual recordings that are freely available online through content sharing websites [17]. The recordings are collected from a wide range of indoor and outdoor locations with varying environmental conditions. The recordings are annotated for whether a bird sound is present at any given time during the 10-second recording. While the challenge is specified as "detection", it can be regarded as "classification" within the terminology of this thesis, since the task is to label the whole 10-second recording instead of detecting the onset and offset. The total length of recordings is around 68 hours [18]. The dataset is publicly available online at http://machine-listening.eecs.qmul.ac.uk/bird-audio-detection-challenge/#downloads.

A. *freefield1010*

A collection of over 7,000 excerpts from **field recordings** around the world, gathered by the FreeSound project, and then standardised for research. This collection is very diverse in location and environment, and for the BAD Challenge we have newly annotated it for the presence/absence of birds. Downloadable separately at https://archive.org/details/ff1010bird

B. *Birdvox-DCASE-20k*

The BirdVox-DCASE-20k dataset contains 20,000 10-second audio recordings. These recordings come from ROBIN autonomous recording units, placed near Ithaca, NY, USA during the fall 2015. They were captured on the night of September 23rd, 2015. Out of these 20,000 recording, 10,017 (50.09%) contain at least one bird vocalization (either song, call, or chatter). The dataset is a derivative work of the BirdVox-full-night dataset [1], containing almost as much data but formatted into ten-second excerpts rather than ten-hour full night recordings. In addition, the BirdVox-DCASE-20k dataset is provided as a development set in the context of the "Bird Audio Detection" challenge, organized by DCASE (Detection and Classification of Acoustic Scenes and Events) and the IEEE Signal Processing Society.

The wav folder contains the recordings as WAV files, sampled at 44,1 kHz, with a single channel (mono). The original sample rate was 24 kHz. Downloadable from https://archive.org/details/BirdVox-DCASE-20k.

C. *warblr10k*

UK bird-sound crowdsourcing research spinout called Warblr. From this initiative we have 10,000 ten-second **smartphone audio recordings** from around the UK. The audio totals around 44 hours duration. The audio will be published by Warblr under a Creative Commons licence. The audio covers a wide distribution of UK locations and environments, and includes weather noise, traffic noise, human speech and even human bird imitations. It is directly representative of the data that is collected from a mobile crowdsourcing initiative. Downloadable separatelay at https://archive.org/details/warblrb10k_public

D. *Chernobyl and PolandNFC*

Chernobyl and PolandNFC datasets were used in task 3 for bird audio detection, namely detecting the presence of any bird in a recording and assigning a file format with the presence of any bird in a recording and assigning a

Fig. 2. 28 most common Malaysian birds.

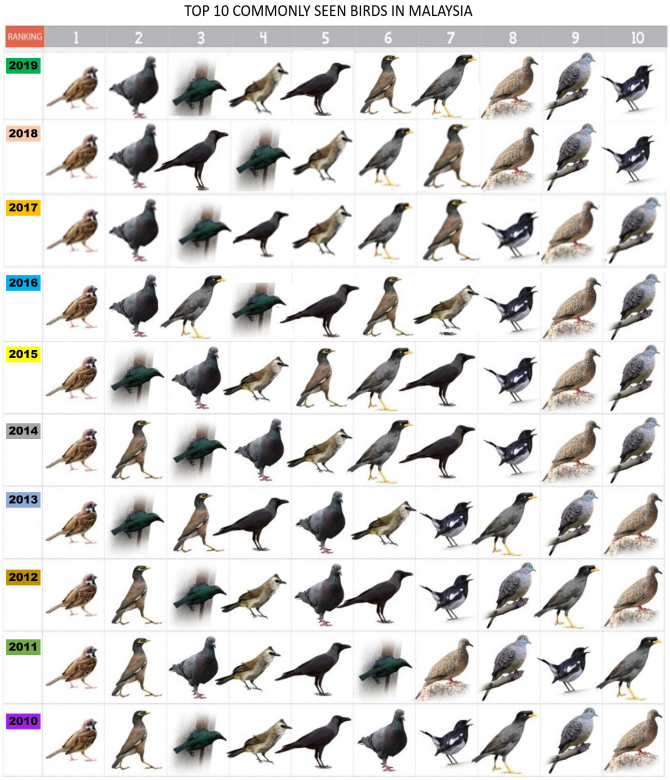


Fig. 3. Ten-year Malaysian bird report.

binary label (1:bird, 0:no-bird) to it (http://dcase.community/challenge2018/task-bird-audio-detection).

The Chernobyl dataset comes from the Natural Environment Research Council (NERC)-funded TREE (Transfer-Exposure-Effects) research project, which is deploying unattended remote monitoring equipment in the Chernobyl Exclusion Zone (CEZ). This academic investigation into the long-term effects of the Chernobyl accident on local ecology is led by Dr Wood. The project has captured approximately 10,000 hours of audio to date (since June 2015). The audio covers a range of birds and includes weather, large mammal and insect noise sampled across various CEZ environments, including abandoned village, grassland and forest areas.

The PolandNFC dataset comes from Pamuła's work for her Ph.D. Both datasets are downloaded through https://archive.org/details/birdaudiodetectionchallenge_test

V. Methodology

A. Proposed Composition of MURAI Dataset

We know that we need at least 800 samples per class, as per Table I. We also know that to test using real-world data, the birds must be resident in Malaysia. We only need 10 species, so the target size of the dataset would around 8000 audio files. The MY Gardens Birds of Malaysia website run by the Malaysian Nature Society collected data on the most common Malaysian birds since 2010 [19]. The most common birds in 2010-2019 are listed Table III. The same 10 species are occupy the top 10 positions every year.

Species selection methodology:

- Start with the most common birds list (28 species)
- Eliminate restricted recordings in Xeno-canto (25 species)
- Choose those with at least 1 hour of total recording (12 species)

B. Dataset construction

Sample preparation:

- Download Xeno-Canto download Python code from https://github.com/ntivirikin/xeno-canto-py
- For each species, download using the Xeno-Canto API. For example, the following downloads quality A for species Zebra dove **foreground** recordings in MP3:

```
# python xenocanto.py -dl Geopelia striata q:A
```

- Only A-grade recordings are considered for now.
- Use Audacity to open the MP3 files one-by-one.
- Mark the start and end of each vocalization in CSV form

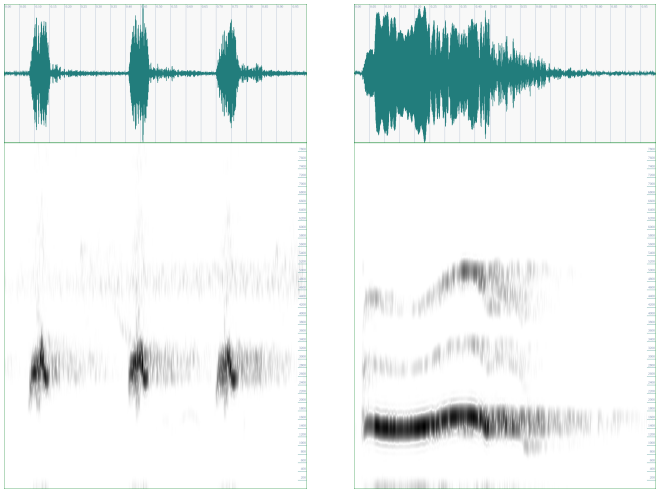
Proposed meta-data in CSV based on Urban8k:

- **slice_file_name**:
  The name of the audio file. The name takes the following format:
  - `[XCID]-[classID]-[occurrenceID]-[sliceID].wav`, where:
  - `[fsID]` = the Xeno-canto ID of the recording from which this excerpt (slice) is taken
  - `[classID]` = a numeric identifier of the sound class (see description of classID below for further details)
  - `[occurrenceID]` = a numeric identifier to distinguish different occurrences of the sound within the original recording
  - `[sliceID]` = a numeric identifier to distinguish different slices taken from the same occurrence

TABLE III
BIRD CONSIDERED FOR MURAI. XENO-CANTO DATA FOR OCTOBER 1, 2020.

English name	Scientific name	q:A # recs	FG # recs	BG # recs	Total # recs	Total Duration	Proposed SpeciesID
Eurasian Tree Sparrow	*Passer montanus*	296	1870	2304	4174	49:11:44	1
Common Myna	*Acridotheres tristis*	87	233	311	544	8:22:46	2
Rock Dove	*Columba livia*	34	147	122	269	1:44:39	3
Spotted Dove	*Spilopelia chinensis*	5	150	302	452	1:23:56	4
Yellow-vented Bulbul	*Pycnonotus goiavier*	7	83	117	200	1:16:10	5
Zebra Dove	*Geopelia striata*	8	112	177	289	1:08:04	6
House Crow	*Corvus splendens*	68	129	313	442	1:03:04	7
Asian Glossy Starling	*Aplonis panayensis*	11	46	17	63	26:39	
Javan Myna	*Acridotheres javanicus*		~~30~~	~~23~~	~~53~~	0:00	†
Oriental Magpie-robin	*Copsychus saularis*		~~402~~	~~87~~	~~489~~	0:00	†
Black-naped Oriole	*Oriolus chinensis*		~~239~~	~~93~~	~~332~~	0:00	†
Asian Koel	*Eudynamys scolopaceus*	125	253	239	492	2:38:31	8
Common Tailorbird	*Orthotomus sutorius*	78	262	148	410	2:08:10	▷
Common Iora	*Aegithina tiphia*	32	214	61	275	1:59:15	9
Olive-backed Sunbird	*Cinnyris jugularis*	48	174	60	234	1:56:14	10
White-breasted Waterhen	*Amaurornis phoenicurus*	22	134	33	167	1:28:23	11
White-throated Kingfisher	*Halcyon smyrnensis*	62	176	56	232	1:24:28	▷
Coppersmith Barbet	*Psilopogon haemacephalus*	9	141	63	204	1:12:17	▷
Ashy Tailorbird	*Orthotomus ruficeps*	51	108	23	131	1:04:45	12
Brown-throated Sunbird	*Anthreptes malacensis*	25	83	9	92	49:56	
Scaly-breasted Munia	*Lonchura punctulata*	4	81	28	109	36:40	
Striated Heron	*Butorides striata*	32	127	20	147	30:53	
Blue-tailed Bee-eater	*Merops philippinus*	21	51	10	61	21:40	
Pacific Swallow	*Hirundo tahitica*	15	44	3	47	19:46	
Blue-throated Bee-eater	*Merops viridis*	7	28	2	30	15:49	
Pink-necked Green Pigeon	*Treron vernans*	6	22	2	24	13:01	
Common Flameback	*Dinopium javanense*	14	31	2	33	12:03	
Pied Triller	*Lalage nigra*	3	25	2	27	7:27	

† Restricted Xeno-Canto, ▷ Not common in East Malaysia.



(a) Yellow-vented bulbul. (b) Asian koel.

Fig. 4. Waveforms dan spectrograms of two common birds.

- **XCID**:
  The Xeno-canto ID of the recording from which this excerpt (slice) is taken
- **start**:
  The start time of the slice in the original Xeno-canto recording, in seconds.
- **length**:
  The length of slice in the original Xeno-canto recording, in seconds.
- **salience**:
  A (subjective) salience rating of the sound. 1 = foreground, 2 = background.
- **fold**:
  The fold number (1-10) to which this file has been allocated.
- **speciesID**:
  A numeric identifier of the species from Table III. Range is 1..12.
- **class**:
  The class name from Table III.

At the moment, the number of species is maintained at 12 based on the availability of raw data of at least 1 hour per species. After the data collection is completed, the top 10 species based the number of slices will be kept.

C. Debugging

For any Xeno-canto recording, it is possible to download using the browser by the URL:

`https://www.xeno-canto.org/<XCID>/download`

where <XCID> is the Xeno-canto recording ID.

Spectrograms can be generated several ways:

- Audacity

- Python librosa code [20]
- Spectrogram online website

D. Utilities

We are making available Python code to read the metadata CSV file and use it to download the source Xeno-canto source audio file and split into the individual vocalizations needed for the dataset. https://github.com/mun3im/murai

REFERENCES

[1] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[2] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv 1804.03209*, 2018.

[3] L. Lai, N. Suda, and V. Chandra, "CMSIS-NN: Efficient neural network kernels for ARM Cortex-M CPUs," *arXiv preprint arXiv:1801.06601*, 2018.

[4] B. McMahan and D. Rao, "Listening to the world improves speech command recognition," *arXiv preprint arXiv:1710.08377*, 2017.

[5] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5484–5488.

[6] J. Lee, T. Kim, J. Park, and J. Nam, "Raw waveform-based audio classification using sample-level CNN architectures," *arXiv preprint arXiv:1712.00866*, 2017.

[7] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] V. Morfi, Y. Bas, H. Pamuła, H. Glotin, and D. Stowell, "NIPS4Bplus: a richly annotated birdsong audio dataset," *PeerJ Computer Science*, vol. 5, p. e223, 2019.

[9] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, p. e488, 2014.

[10] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Birdvox-full-night: A dataset and benchmark for avian flight call detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 266–270.

[11] J. G. Arriaga, M. L. Cody, E. E. Vallejo, and C. E. Taylor, "Bird-DB: A database for annotated bird song sequences," *Ecological Informatics*, vol. 27, pp. 21–25, 2015.

[12] J. Salamon, J. P. Bello, A. Farnsworth, M. Robbins, S. Keen, H. Klinck, and S. Kelling, "Towards the automatic classification of avian flight calls for bioacoustic monitoring," *PlOS ONE*, vol. 11, no. 11, pp. 1–26, 2016.

[13] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[14] H. Pamuła, M. Kłaczyński, M. Remisiewicz, W. Wszołek, and D. Stowell, "Adaptation of deep learning methods to nocturnal bird audio monitoring," in *Postępy akustyki*. Polskie Towarzystwo Akustyczne, Oddział Górnośląski: Wydawnictwo Infoart, 2017, pp. 149–158.

[15] E. Scholes III, "Macaulay library audio and video collection," 2017.

[16] K.-H. Frommolt, "The archive of animal sounds at the Humboldt-University of Berlin," *Bioacoustics*, vol. 6, no. 4, pp. 293–296, 1996.

[17] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: a survey and a challenge," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

[18] E. Cakir, "Deep neural networks for sound event detection," Ph.D. dissertation, Tampere University of Technology, 2019.

[19] Malaysian Nature Society, "MY Garden Birdwatch." [Online]. Available: https://www.mygardenbirdwatch.com/

[20] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proceedings of the 14th Python in science conference*, vol. 8, 2015, pp. 18–25.