[Bird Song Identification]

[Gold Team]

Data Science Capstone Project
Data Acquisition and Pre-Processing Report

Date:

[05/02/2024]

Team Members:

Name: Jonathan Watkins

Name: Joseph Trybala

Name: Max Song

Identifying Data

Data Sources:

We procured our data from *xeno-canto* via Kaggle(Rao, n.d.). The decision to utilize these data sources for our project was based on several criteria, which encompasses both the limitations and capstone project objectives stipulated by DSCI591 itself, as well as the acknowledged time constraints of a Drexel quarter. Among multiple esteemed data repositories for bird calls and birdsong, we selected the aforementioned, Kaggle and xeno-canto. Kaggle is recognized for their extensive, open-sourced repositories of data and datasets for the purpose of data analysis and machine learning. This choice underscores our commitment to employing data obtained from reputable sources as well as maintaining the integrity of the data used within the context and for the purpose of completing our capstone project.

Kaggle hosts the dataset(s), two links which reflect a singular dataset containing recordings of 264 different species along with the corresponding metadata (species data, contributor, length of recording, bit rate, lat/long coordinates, self-report recording location, date of recording, etc.) (This metadata can be found attached to the recordings themselves) which have been broken down into two subsets, one containing bird species A-M (18GB in total) and the other species with starting alphabet N-Z (11GB in total). Per Kaggle, as each recording is its own entity, the license for each can be found in a metadata file (*train_extended.csv)*. This dataset is updated on a weekly basis. The dataset has also been used yearly as part of the BirdCLEF competition hosted through Kaggle, which often focuses on under-studied bird species with the goal of enrichment of the dataset itself (Kahl et al., 2023).

Further, the site www.xeno-canto.org, referred to as “Xeno-canto” or simply “XC” is itself a website created for the purpose of sharing wildlife sounds. The project was undertaken by Robert Planqué and Willem-Pier Vellinga. The site itself is maintained by a team of administrators in cooperation with *Naturalis Biodiversity Center*. This center is both a national museum of history and research center for the Netherlands located in Leiden in Europe. The website itself is officially run by the Xeno-canto foundation, *Stichting Xeno-canto voor natuurdluiden* from the Netherlands.

Acquisition Process:

The data has been acquired through Kaggle. The data is hosted publicly through the two links associated with the dataset. As mentioned in the previous section, these form one complete dataset. All of the data can be downloaded via the Kaggle website on any standard web-browser. Each link is available for download as a zip file which contains the recordings. This format ensures the data is easily retrievable and manageable, particularly given the substantial file sizes totaling 29GB across both subsets. The data is sourced from www.xeno-canto.org as stated. Both entities fall within publicly accessible domains (Rao, n.d.).

The data acquisition process for our project is managed through use of Kaggle, offering a direct and user-friendly method for accessing the datasets. The dataset is made publicly available and accessible via the two distinct links on Kaggle. These links form the unified dataset, as detailed in the previous section and facilitate a straightforward approach to data procurement.

While this data is already segmented into two parts for ease of accessibility, i.e. download, no additional integration is needed. This allows for us, specifically, but also other researchers and teams to focus on the analytical goals without the necessity of data merging. Moreover, as such, the data and dataset requires no custom code for acquisition. No further methods were required for data retrieval.

Data Acquisition Issues:

At the time of submission of this report there are no known issues with data acquisition. It should simply once again be mentioned that the original dataset is updated on a weekly basis and it is possible that organically, it may undergo amendments, subtractions, or other alterations, as it is not explicitly stated what 'updates' translates to in terms of changes to the data itself. Because we are using a packaged version found on Kaggle, the data used is static and will be the same for all users provided the data is downloaded from the same source.

Data-Processing

[Examine the data you have acquired and understand the data properties. Is there any pre-processing you need to do before you can start analyzing the data? For example, missing data, sparsity, noise, veracity, ambiguity, interoperability, etc. Describe each data issue in a sub-section and explain how you clean up the data.]

No extensive preprocessing is required. Only data type conversion is needed for the approach we have chosen. Instead of working with pure sound data we are converting to spectrograms. We have checked the data over, and no preprocessing should be necessary. If we were to have acquired the same data from elsewhere then preprocessing would have been vital, this led to our search for a better package of the same data which is this prepulled Kaggle set. To elaborate: when this birdsong identification competition was declared Cornell referred everyone to the Xeno Canto public API to pull and collect the data assembled on the site by volunteers. This is a weighty task, and there were citable issues with the API because of the size of the data involved (as well as the amount of people giving a go at the competition). The package we have chosen was pre-pulled and organized into logical buckets for convenience. We checked for any additional alteration of the data, and detected none.

For first step processing we need to convert this audio data to visual data in the form of spectrograms. This processing pipeline has already been built and tested and uses the librosa library. A code sample is available in the attached .ipynb file. Additional processing includes alteration of spectrograms after creation, this is done to suppress noise and boost signal, as well as segment or chunk the data into forms better suited for training.

Missing Data

Upon thorough screening of the dataset sourced from xeno-canto via Kaggle, it is evident that there are no significant issues with missing data. Every bird species within our dataset has at least one corresponding audio file. This comprehensive coverage ensures that no crucial information is absent due to missing data. Consequently, there is no need for imputation or deletion of incomplete records, as the dataset is complete in terms of audio samples for each bird species.

Sparsity

The dataset exhibits significant sparsity in terms of the distribution of bird species and their respective songs. Unlike an evenly distributed dataset where each category or class has a comparable number of samples, our dataset shows considerable variation in the number of audio samples per bird species. For instance, bird species like the Black-chinned Hummingbird and Black-billed Cuckoo are represented by only a few samples (1 and 4, respectively), while other species such as the Black-capped Chickadee may have more than 100 audio samples. This imbalance in sample distribution across different bird species poses a challenge for analysis and may require careful consideration during model training and evaluation to avoid biases towards overrepresented species. We do not intend to prune underrepresented species unless model performance is deemed poor.

Noise

In our dataset, we confront the pervasive challenge of noise interference currently present in the field recordings. Our investigation into noise mitigation strategies stems from an extensive literature review paper that systematically explored various methodologies for bird song identification within the existing literature.

Noise within recordings obscures the clarity of bird calls, impacting the performance of recognition algorithms. Leveraging insights from our literature review, we adopt a multifaceted approach to address this issue and enhance signal quality (Wildlife Acoustics 2011) and Boucher 2014)). By integrating recordings with minimal noise contamination into our training data, inspired by findings from prior research, we aim to minimize false positives during identification tasks. Furthermore, we may select song examples for our future model from environments akin to our test recordings to ensure consistent noise profiles and reduced call variations (Katz et al. 2016),.

Recognizing the critical role of advanced denoising techniques, as emphasized by Schrama et al. (2007), we explore innovative methods identified through our literature review. Additionally, drawing from the work of Priyadarshani et al. (2016) and Priyadarshani (2017), we may also investigate wavelet-based denoising methods optimized for our automatic sound recording setup, addressing stationary background noise, including geophony (wind, rain, etc.), across various bandwidths. Moreover, inspired by the insights presented by Potamitis (2014), we are exploring the potential of treating spectrograms as images and applying image analysis and computer vision techniques to mitigate noise. While this approach aids in detecting regions of interest, we remain cognizant of its limitations in terms of reversibility and general utility for identification tasks within our project.

While not pertinent to data acquisition there are several obstacles that birdsong identification via field recordings can have, that will directly impact our work and must be accounted for, (Priyadarshani et al., 2018) which we divide into general and preprocessing issues:

General Issues

We identify these as fundamental challenges that are in inherent in recording(s) and analyzing birdsong, but not necessarily linked to actual more-technical preprocessing steps

1. *Environmental Noise Overlap*: Included herein are all unavoidable sounds from weather, other animals, and includes man-made sources such as transportation, i.e. planes and perhaps machinery such as wind turbines that may interfere with the clarity of the birdsong recordings.
2. *Variability Within Species*: It is to be acknowledged that birdsong can vary immensely within a species due to repertoire size and complexity, due to geographical variations, and ultimately the ability of some bird species to innovate new songs.
3. *Adaptation to Environment:* Birds may adapt their songs in response to environmental stimuli, including man-made noise (referred to as spectral modulation), and may alter the duration and composition of their songs based on differing situations (temporal, location).
4. *Juvenile Variations*: Young birds have a tendency to produce unusual calls while learning their species' typical songs, similar to human baby babble, leading to variability in recordings of juvenile versus adult birds.

Preprocessing Issues

These issues are directly related to the technical aspects of handling and preparing the audio data for analysis:

1. *Variable Levels of Power(dB)*: Differences in recording volume due to the bird's distance from the recorder or its general orientation relative to the microphone can affect the consistency and overall quality of the audio data.
2. *Song Duration Variations*: Birdsong can differ in length and be changes which may occur due to situational adjustments by the birds, further complicating the preprocessing stage where consistent song lengths might be necessary for effective analysis.

Appendix

[Provide the code or pseudo code, data definition, sample data, and any other information in the appendix here.]

Code is supplied as an attached .ipynb notebook.

Below is an example spectrogram of an audio sample of an adult Alder Flycatcher. If you look at the range above 2000 hertz you can see a call in between 5 and 10 seconds. This is considered a good representation of a call and we will be applying filters to further promote the calls prominence vs background noise.

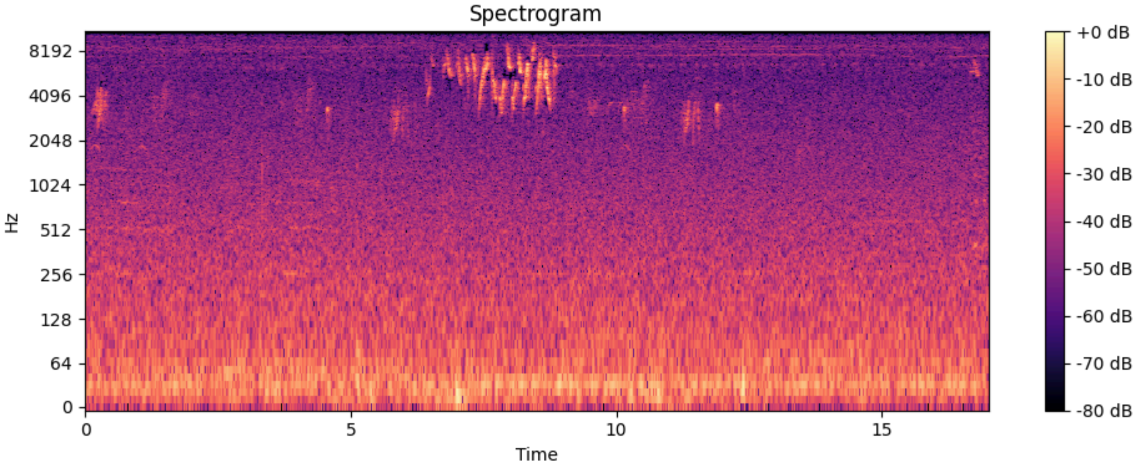

Table of Contributions

The table below identifies contributors to various sections of this document.

	Section	Writing	Editing
1	**Data Sources**	**J.W., M.S**	**All**
2	**Data Pre-Processing**	**All**	**All**
3	**Appendix**	**J.T.**	**All**

Grading

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.

References

Boucher, N. J. (2014). SoundID version 2.0. 0 documentation.

Kahl, S., Denton, T., Klinck, H., Reers, H., Cherutich, F., Glotin, H., Goëau, H., Vellinga, W.-P., Planqué, R., & Joly, A. (2023). Overview of BirdCLEF 2023: Automated Bird Species Identification in Eastern Africa. In *Proceedings of the International Conference on Multimedia Retrieval*.

Katz, J., Hafner, S. D., & Donovan, T. (2016). Assessment of error rates in acoustic monitoring with the R Package Monitor. *Bioacoustics*, *25*(2), 177–196. https://doi.org/10.1080/09524622.2015.1133320

Potamitis, I. (2014). Automatic classification of a taxon-rich community recorded in the wild. *PloS one*, *9*(5), e96936.

Priyadarshani, N., Marsland, S., Castro, I., & Punchihewa, A. (2016). Birdsong denoising using wavelets. *PloS one*, *11*(1), e0146790.

Priyadarshani, N. (2017). *Wavelet-based Birdsong Recognition for Conservation : A thesis presented in partial fulfillment of the requirements for the degree of doctor of philosophy in computer science at Massey University, Palmerston North, New Zealand*. Massey Research Online. http://hdl.handle.net/10179/12127

Priyadarshani, N., Marsland, S., & Castro, I. (2018, January 6). *Automated birdsong recognition in complex acoustic environments: a review*. Nordic Society Oikos. https://doi.org/10.1111/jav.01447

Rao, R. (n.d.). Xeno-canto bird recordings extended [A-M]. Retrieved [April 5th, 2024], from https://www.kaggle.com/datasets/rohanrao/xeno-canto-bird-recordings-extended-a-m

Rao, R. (n.d.). Xeno-canto bird recordings extended [N-Z]. Retrieved [April 5th, 2024], from https://www.kaggle.com/datasets/rohanrao/xeno-canto-bird-recordings-extended-n-z

Schrama, T., Poot, M., Robb, M., & Slabbekoorn, H. (2007, December). Automated monitoring of avian flight calls during nocturnal migration. In *International Expert meeting on IT-based detection of bioacoustical patterns* (pp. 131-134).

Wildlife Acoustics (2011). *Song Scope bioacoustics software version 4.0 documentation.* – Wildlife Acoustics.