Audio Segmentation to Build Bird Training Datasets

Diego T. Terasaka[1], Luiz E. Martins[1], Virginia A. dos Santos[1], Thiago M. Ventura[1], Allan G. de Oliveira[1], Gabriel de S. G. Pedroso[1]

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
Cuiabá – MT – Brasil

diego.takao.t@gmail.com, luiz.martins@sou.ufmt.br, lownialore@gmail.com,
thiago@ic.ufmt.br, allan@ic.ufmt.br, gabriel.de.s.g.pedroso@gmail.com

Abstract. *To create a bird classification model, it is necessary to have training datasets with thousands of samples. Automating this task is possible, but the first step is being able to segment soundscapes by identifying bird vocalizations. In this study, we address this issue by testing four methods for audio segmentation, the Librosa Library, Few-Shot Learning technique: the BirdNET Framework, and a Bird Classification Model called Perch. The results show that the best method for the purpose of this work was BirdNET, achieving the highest values for precision, accuracy, and F1-score.*

1. Introduction

The field of bioacoustics serves as a crucial tool for monitoring animal vocalizations in wildlife, such as birds, facilitating the detection of changes within ecosystems. Artificial intelligence models contribute significantly to this effort. However, it is necessary to use datasets containing thousands of samples for each species of birds and to utilize labeled data to enhance the efficiency of the samples and improve overall data quality (Chen et al., 2008).

Our long-term goal is to automatically prepare sets with similar acoustic characteristics to enable the construction of training datasets for classifying bird species, specifically from Brazilian Pantanal. To achieve this, the following stages must be developed: select audio snippets containing bird vocalizations from soundscapes; extract characteristics from selected snippets; group snippets with similar acoustic characteristics; evaluate the groups by assigning a bird species; and validate the training datasets for each assigned species.

In this work, we detail the first stage of this project, which consists of evaluating and selecting the best audio segmentation method for detecting vocalizations of bird species. Next, we describe four evaluated methods, the datasets used, and the results achieved to fulfill the objective of this stage of the project.

2. Materials and Methods

Some works have been done for this purpose, such as García-Ordás et al. (2023) and Xue et al. (2023). Based on these ideas, four audio segmentation methods were evaluated, considering code availability: (i) audio segmentation using Librosa library (McFee et al., 2015); (ii) a method implemented based on Few-shot Bioacoustic Event Detection (Nolasco et al., 2023); (iii) using BirdNET model (Kahl et al., 2021); (iv) another bird classification model named Perch (Google Research, 2023).

The first method, named “Frequency Detection”, consists of two main functions or stages. Firstly, the initial function aims to reduce noise in the audio signal using Short-Time Fourier Transform (STFT). Subsequently, the processed audio is passed to the second function to generate a vector of arrays containing integers, where each element or array represents the start and end times of the events of interest. This process involves analyzing the audio spectrum to identify regions with the desired frequencies, resulting in a spectrogram-to-time transformation and ultimately yielding a track of the desired events.

The method named “Few-shot Bioacoustic Event Detection” aims primarily to be used for long-duration recordings, with the premise of using examples where the bird sings, and then mapping the rest of the audio accordingly. For these examples, audio recordings from Xeno-Canto 2024 were selected, which is a collaborative online platform dedicated to sharing bird vocalization recordings from around the world. In order to test the algorithm and assess how it performed with different sound frequencies, five audio recordings of the main birds of the Pantanal Mato-Grossense were selected, such as the *Herpetotheres Cachinnans, Ortalis Canicollis, Cercomacra Melanaria, Cyanocorax cyanomelas, Crypturellus Parvirostris* and *Synallaxis Albilora*.

BirdNET and Perch are both convolutional neural network models based on the EfficientNet (Tan & Le, 2019). BirdNET is currently updated to its 2.4 version, with code available online along with supporting interfacing scripts. Although these neural network models pertain to the classification of bird species, an appropriate parsing of their analysis’ results should be able to indicate generic bird activity when considering the lower confidence results.

After code implementation, we evaluated each method according to its ability to detect bird vocalizations. For this comparison, fifteen audio samples were used having bird vocalizations, ambient noises such as rain, and human songs, as well as silent recordings. The data used for processing, from Ventura et al. (2024), was obtained in the Pantanal Mato-Grossense, Brazil. These recordings were captured during two different time periods: dry and rainy seasons. The samples consisted of 5,734 seconds of recording, with individual recordings ranging from a minimum of 180 seconds to a maximum of 900 seconds. The onset and offset times of bird vocalizations were manually identified and recorded by three researchers.

Accuracy, Precision, Recall, and F1 metrics were collected to evaluate the performance of the methods in order to choose the best one for our purpose.

3. Results and Discussions

All files described in the previous section were processed with the four selected methods. The estimates generated by the models were compared with the intervals manually assigned by the researchers. The performance of each of them can be seen in Figure 1.

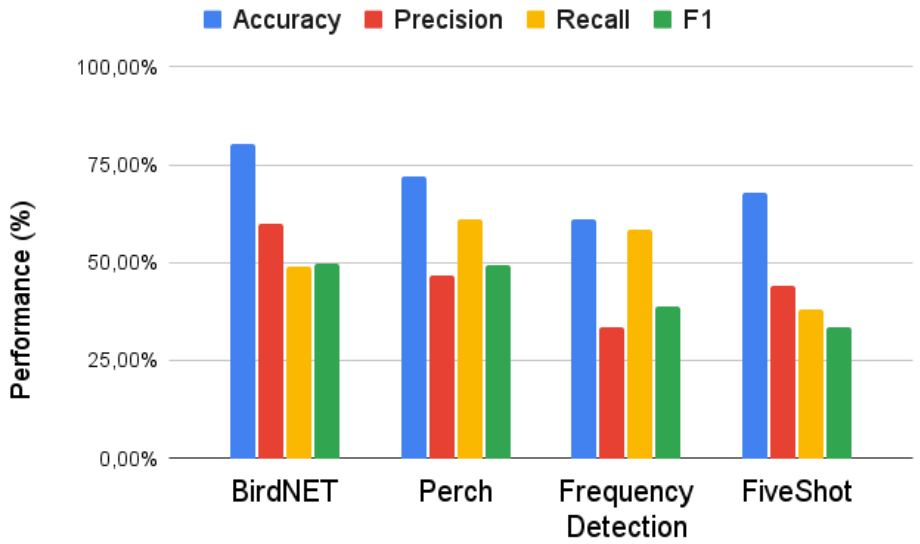

Figure 1. Performance of the methods considering the average of each metric evaluated

The neural network methods have shown better overall performance, with each excelling at different metrics while achieving similar F1 scores (BirdNET 50%, Perch 49%). BirdNET has shown less propensity to garnering false positive predictions, achieving higher precision (60%) and lesser recall (49%) than the other methods. Perch amasses a higher positive flag count, greatly reducing false negative count while attaining more false positive results, defining a higher recall (61%) and a lower precision (47%). The Frequency Detection approach, although scoring the lesser precision (33%) and F1 (39%) of the three, still achieves a greater recall (59%) than BirdNET, excelled only by Perch. Although the "FiveShot" method achieved higher precision and accuracy compared to Frequency Detection, with accuracies of 68% and precision of 44%, respectively, "Frequency Detection" obtained a higher recall, reaching 59% recall and 61% accuracy. Conversely, the FiveShot method exhibited lower recall and F1 scores compared to the other three methods, with recall at 38% and F1 at 34%. This model appears to have a considerable number of false positives, but when it makes correct predictions, it demonstrates good precision.

Even though there are works with high performance, such as that of Narasimhan et al. (2017) with a true positive rate of 98%, the values found in this work guarantee the achievement of our objective, which is to confidently select different vocalizations to compose training bases, without the need to detect all existing vocalizations. The requirement for trustworthy vocalization snippets prioritize precise predictions over better encompassing options. For some reason the metric of precision must be prioritized when choosing the best method for our purpose. Therefore, the use of BirdNET characterizes a more ideal approach for the current objective.

4. Conclusions

As mentioned, thousands of examples are needed to compose training datasets to build models for applications in bioacoustics. This process takes time and resources, making it challenging to classify hundreds of different species, such as birds. This task can be automated by completing several steps, the first of which is audio segmentation when a bird vocalization is identified, as addressed in this work.

We investigated methods for performing audio segmentation and tested four of them. The BirdNET (Kahl et al., 2021), a classifier model capable of identifying 984 bird species, obtained the best results for identifying different bird vocalizations.

The next step is to apply this method to several soundscape recordings, obtaining thousands of snippets of bird vocalizations. Next, features of these snippets will be extracted and grouped, with the expectation that audios of the same bird species will be indicated in the same group. Each group will represent a species and their audios can form a new training dataset, achieving our long-term goal.

References

Chen, H. L., Chuang K. T., and Chen M. S., (2008) On Data Labeling for Clustering Categorical Data. IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 11, pp. 1458-1472, Nov. 2008, doi: 10.1109/TKDE.2008.81.2

García-Ordás, M. T., Rubio-Martín, S., Benítez-Andrades, J. A., et al. (2023). Multispecies bird sound recognition using a fully convolutional neural network. Applied Intelligence, 53, 23287–23300.

Google Research (2023). Google Bird Vocalization Classifier: A global bird embedding and classification model. https://tfhub.dev/google/bird-vocalization-classifier/4.

Han, X., & Peng, J. (2023). Bird sound classification based on ECOC-SVM. Applied Acoustics, Volume 204, 2023, 109245.

Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics, 61, 101236.

McFee, B., Raffel, C., Liang, D., Ellis, D. PW., McVicar, M., Battenberg, E., Nieto, O. (2015) Librosa: Audio and music signal analysis in python. Proceedings of the 14th python in science conference, pp. 18-25.

Narasimhan, R., Fern, X. Z, Raich, R. (2017). Simultaneous segmentation and classification of bird song using CNN. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 146-150, doi: 10.1109/ICASSP.2017.7952135.

Nolasco, I., Singh, S., Morfi, V., Lostanlen, V., Strandburg-Peshkin, A., Vidaña-Vila, E., Gill, L., Pamuła, H., Whitehead, H., Kiskin, I., Jensen, F. H., Morford, J., Emmerson, M. G., Versace, E., Grout, E., Liu, H., Ghani, B., & Stowell, D. (Eds.). (2023). Learning to Detect an Animal Sound from Five Examples. Ecological Informatics. 77.

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. International conference on machine learning (pp. 6105-6114).

Ventura T. M., Ganchev, T. D., Granados, C. P., Oliveira, A. G., Pedroso, G. S. G., Marques, M. I. and Schuchmann K. L. (2024) The importance of acoustic background modelling in CNN-based detection of the neotropical White-lored Spinetail (Aves, Passeriformes, Furnaridae). Bioacoustics, DOI: 10.1080/09524622.2024.2309362

Wang, H., Xu, Y., Yu, Y., Lin, Y., & Ran, J. (2022). An Efficient Model for a Vast Number of Bird Species Identification Based on Acoustic Features. Animals, 12(18), 2434.

Xeno-Canto. Sharing wildlife sounds from around the world. 2022. [accessed 2024 March 06]. https://xeno-canto.org.