Audio Segmentation of Kauai Forest Bird Vocalizations Using Logistic Regression

Ivy Wang
wangivy@stanford.edu

Summary

The task of determining whether bird species are present in pre-segmented audio recording has been a well-studied area in machine learning. Conservationists from the Kauai Forest Bird Recovery Project hope to apply these algorithms to field recordings in order to determine bird counts from field recordings, but the important pre-processing step of segmenting the audio has been under-studied and is currently done manually.

The goal of this project is to use logistic regression to automate the task of segmenting audio from field recordings into relevant sections where bird vocalizations (calls and chirps) are present. The results of this project show that logistic regression performs as well as state-of-the-art techniques for bird vocalization segmentation tasks. However, all current techniques remain bottle-necked by the quality and quantity of the data available to train these models.

Data

Field recordings[1] in the form of .wav files were obtained from Kauai Forest Bird Recovery Project, and were preprocessed by running noise reduction, then manually labeled using timestamps corresponding to the beginning and ending times of audio segments which contained bird vocalizations. The positive labels were sparse, and logistic regression is sensitive to imbalanced data, so the data was balanced by arbitrarily discarding sections of the audio where no vocalizations were present. Furthermore, the labeling process was imperfect, so the resulting ground truth labels were imprecise and contained erroneous labels.

Features

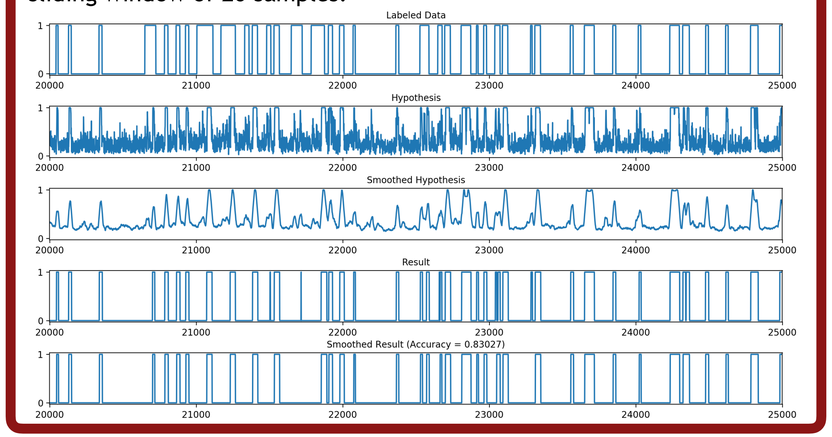
The audio was segmented using a sliding window of length 25 ms and a step size of 10 ms. Mel-frequency cepstral coefficient (MFCC) features were then extracted for each segment. These features were chosen because the frequency bands used by MFCC approximates the response of the human auditory system and are commonly used for similar audio recognition tasks[2,3]. Other commonly used features such as energy were not chosen because the task required identifying bird sounds in a noisy environment where a bird vocalization did not necessarily correspond to such features.

Model

The parameters (θ) of the model are applied to form a (scalar) linear combination of the 26 MFCC features, which is then passed into the sigmoid function to give a hypothesis (corresponding to the likelihood that a window contains vocalizations). These parameters are tuned using stochastic descent by minimizing the following cost function:

$$J(\theta) = \sum_{i=1}^{m} y_i \log(\sigma(X_i\theta^T)) + (1 - y_i)\log(1 - \sigma(X_i\theta^T)) + \frac{\lambda}{2m}|\theta|^2$$

The resulting θ is then applied to obtain a hypothesis, which is smoothed via a sliding window of 20 samples. The result is obtained by marking each window as positive iff the hypothesis is above the threshold, and then smoothed using a majority vote protocol across a sliding window of 20 samples.



Results

The performance of the data was measured by checking whether the categorization of each window matched the ground truth label of the corresponding timestamp. Note that the data had very high variance, which also resulted in errors with very high variance depending on how the data was partitioned (i.e., the chosen sets did not fully represent the range of vocalizations present, so more data would have given more meaningful results).

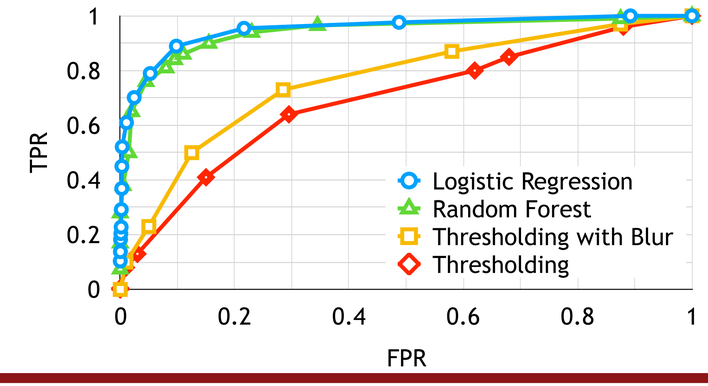
Data	Number of Samples	Error
Balanced Training Set	16,233	19.0%
Balanced Validation Set	5,411	18.7%
Balanced Test Set	5,411	9.8%
Imbalanced Test Set	160,800	17.0%

Discussion

With a threshold of c = 0.3, the model captured 88.9% of the bird vocalizations with a false positive rate of 8.6%. This is on par with state-of-the-art techniques[3], and significantly better than frequency and energy thresholding techniques[4] used in the past.

The calculated accuracies are an underestimate of performance due to the imprecise nature of the ground truth labels. For example, bird vocalizations where there are multiple chirps are manually labeled (incorrectly) as a single segment of bird vocalizations, whereas the algorithm labels each chirp separately and does not mark the intervals between the chirps as positive. These intervals are considered false positives, despite the algorithm outperforming human precision. This suggests that the performance of algorithms on such tasks are currently being limited by the quality of data available, and more performant algorithms may be difficult to obtain without improving data quality first.

ROC Curve



Future Work

Given that the algorithms are currently bottlenecked by the quality and quantity of data labels, a necessary first step would be to obtain more labeled data of better quality. Furthermore, the current model only considers linear relationships between the MFCC features. One area to explore would be to augment the feature space by including non-linear features.

[1] https://tinyurl.com/229a-bird-noise
[2] Xu, Min, et al. "HMM-based audio keyword generation." *Pacific-Rim Conference on Multimedia*. Springer, Berlin, Heidelberg, 2004.
[3] Neal, Lawrence, et al. "Time-frequency segmentation of bird song in noisy acoustic environments." *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011.
[4]P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 2252 – 2263, November 2006.