UNIVERSITY OF ALBERTA

FACULTY OF SCIENCE
DEPARTMENT OF BIOLOGICAL SCIENCES

CW 405, Biological Sciences Bldg.
Edmonton, Alberta, Canada T6G 2E9
[redacted]

[redacted]

Thank you for your consideration of our manuscript, “Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs”. Like the original submission, the results in this manuscript have been submitted exclusively to *Avian Conservation and Ecology*. My co-authors are aware of, and have approved the submisison of this revised version.

Our sincere thanks to the Subject Editor and three Reviewers, whose thoughtful comments and broad perspectives greatly improved the accuracy, message, and utility of our manuscript. We considered every comment and addressed each one while striving to retain language and meaning accessible by ornithological practitioners. The following documents both the major and minor changes we made to the manuscript. We have refrained from detailing responses to comments on writing style and clarity, but we made use of the vast majority of the suggestions made by the Reviewers. We would be happy to compile a full concordance table if you require.

Sincerely,

Elly C. Knight, Ph.D. Student

UNIVERSITY OF ALBERTA

FACULTY OF SCIENCE
DEPARTMENT OF BIOLOGICAL SCIENCES

CW 405, Biological Sciences Bldg.
Edmonton, Alberta, Canada T6G 2E9

MAJOR REVISIONS

1. **Balanced training data:** We have balanced the training data across all recognizers, reprocessed our test data, and rerun all the analyses. We added the remaining 50 training clips to the MonitoR recognizer, and as expected, it doubled the scanning time, but did not substantially improve the performance of the recognizer. We also reduced the number of true negative training clips in the Kaleidoscope recognizer to 100. The recognizer performance has increased, but this is due to other improvements made in response to a comment by Reviewer C. Please see Major Revision #7 for details. All results and conclusions remain the same after balancing the training data.
2. **Efficiency units and implications:** We agree with Reviewer C and the Subject Editor that processing time should be reported in hours of audio recording and not GB, and have made this change. We have also changed Figure 7 to a table and included the dataset size at which recognizer processing becomes faster than human listening instead of a range of dataset sizes because it provides a more detailed description of relative efficiency than the figure. Finally, we have added more detail to the description of the efficiency calculations in the methods for transparency.
3. **Inclusion of literature:** We have now cited most of the literature suggested by Reviewer B; however, we have refrained from an extensive expansion of literature cited to maintain the accessibility of our MS for ecologists. The signal recognition literature is vast and highly technical. Much of this literature is also not directly relevant to our work, including some of the papers suggested by Reviewer B, which deal with multiclass recognizers. These are not relevant to our MS on single class recognizers. We have included the following references:
    - Expanded the section on state-of-the art approaches in the introduction, including a reference to the BirdCLEF challenge and other multiclass bird recognition challenges and additional citations of deep machine learning methods.
    - Added references to papers that use convolutional neural networks.
    - Added details throughout the MS on the existence of established best practices for signal recognizer evaluation in other fields, using the papers suggested by Reviewer B. See Major Revision #4 for further details.
4. **Scope of recommendations:** We thank Reviewer B for pointing out our poor communication of the notion of the novelty of our recommendations. Our intention was not to claim these recommendations as novel, but to synthesize them in a manner accessible to ecologists, as the lack of assessment literacy demonstrated in our review of the

ecological literature suggested there is a need for general recommendations in ecology. To that end, we have added the following:
    - Specified in the introduction that our work is intended for ecological practitioners and not state-of-the-art bioacoustic specialists
    - Noted that we are drawing our recommendations from best practices in other fields throughout the paper. We have added this to the abstract, last paragraph of the introduction, methods (beginning and metrics section), recommendations (beginning and metrics section), and discussion.
    - Added the other best practices papers suggested by Reviewer B.
5. **Occupancy analysis:** We have changed and improved the occupancy analysis in response to concerns from Reviewer A and C because occupancy modelling is an important example of application evaluation for ecologists. Instead of running single models with score threshold as a covariate, we have run separate null occupancy models for each of the 0.01 score threshold bins and simply plotted the mean and 95% confidence intervals for each model. The results suggest that occupancy estimates from recognizer data are particularly problematic when the recognizer has low recall. We feel this is an important preliminary result on the topic of using recognizer data for occupancy modelling and have recommended further research in the discussion, as suggested by Reviewer A.
6. **Generalizability:** The training and test data for the performance evaluation were in fact from two different geographic locations (training from BC and testing from ON) as a way of demonstrating generalizability. We have now specified this explicitly in the methods of the manuscript. At the suggestion of the Subject Editor, we have also added generalizability as a short additional recommendation.
7. **Kaleidoscope recognizer:** The 'perplexing' results of our Kaleidoscope recognizer and reviewer comments on learning time investment encouraged us to revisit our results. We now include results that are more congruent with the other recognizer programs. In particular, Reviewer C's suggestion to limit number of clusters to 2 made the direction of metric responses relative to score more logical, although the trend is still erratic in some cases, likely due to the clustering nature of this program. The recent release of a user manual for Kaleidoscope also allowed us to understand and revisit the other clustering and signal detection parameters. We set maximum cluster distances to simulate the other recognizers as much as possible, and use the same signal detection settings as the Song Scope recognizer. The end result was a recognizer that performs much better than our first attempt, but remains only moderately effective compared to the other programs tested.

MINOR REVISIONS & RESPONSES

1. **Terminology:** We changed "assessment" to "evaluation" throughout the manuscript because "assessment" implies testing to recommend improvements, as opposed to "evaluation", which implies testing to judge functionality.

2. **F-score**: Given the suggestion of Reviewer B, we have included F-score in our recommended evaluation metrics. We had initially left it out because the β value of the F-score renders this metric subjective to user priority of recall versus precision; however, in light of Reviewer B's emphasis on the importance of this metric, we included it with the recommendation that F-score be reported with $\beta = 1$ to allow for comparison across papers.
3. **Human efficiency:** We have added a caveat about the efficiency of scanning relative to human listening in Recommendation 5 (Efficiency Evaluation). We also added a note about the efficiency of multispecies data processing within this recommendation.
4. **Appendix 2:** We have included more references to the appendix in the text to point the reader towards the details requested by Reviewer B (full CNN architecture and minimum score thresholds used).
5. **State-of-the-art:** While we agree with Reviewer B that the programs tested in the manuscript are not state-of-the-art, we included them because these are the programs currently available to and used by ecologists without bioacoustic or machine learning expertise. We therefore argue the comparison of these programs is valuable for ecologists, as the average practitioner is not capable of employing the current state-of-the-art. We have also clarified the intent of our statement within the introduction that there are no existing comprehensive comparisons of recognizers to specify 'commercially and freely available recognizer programs'.
6. **Top hits for moving window recognizers:** In response to comments from Reviewer B and C, we would like to expand on our rationale for the choice of using the top 6,750 hits for the moving window recognizers. We chose to limit the number of hits from the moving window recognizers in an effort to reduce the bias of this type of recognizer because it will always produce more hits than a signal detection recognizer if run with a score threshold of 0; however, many of those hits will not be actual signals because there is no signal detection process, and thus will be low-scored false positives. To reduce the occurrence of this, we chose the maximum number of hits by any of the signal detection recognizers (6,750) as the maximum allowable hit limit for the moving window recognizers. This limit was chosen based on the assumption that 6,750 hits was the number of signals in the test dataset and that the moving window recognizers would detect those 6,750 signals first before reporting non-signal hits.
7. **Focus on score threshold:** We have chosen to place heavy emphasis on score threshold in our MS given our literature review. It is not common practice in ecology to include all score thresholds from 0 to 1 in recognizer evaluation. Although Reviewer B argues that the practice of inclusion of all score thresholds is an obvious best practice, this has not been taken to heart by ecologists; therefore, it is important to stress the fundamental importance of score to a paper geared towards our target audience.
8. **Effective detection radius:** We argue in the discussion that some of the discrepancy between human listening and recognizer processing is due to a difference in effective detection radius. We maintain this argument in response to comments by Reviewer C. We

have not detailed this argument in detail in the manuscript for the sake of brevity, but we are currently preparing another MS that shows score is actually a proxy for detection distance if the recognizer is trained with clips of calls close to the ARU, which our recognizers were. The recognizer is therefore searching for calls near the ARU and so misses some of the further calls because the acoustic signal deviates from what it is trained to search for as sound attenuates with distance. In contrast, the human observer implicitly knows what a call sounds like at near, mid-range and far distances, and so the decline in detectability with distances is not as steep as for the recognizer. The sensitivity of the recognizer's signal detection process may also contribute to smaller effective detection radius if the program is less proficient at discriminating acoustic signal from background noise than a human.

9. **Spectrogram figure:** We have added a spectrogram of the Common Nighthawk call as suggested by Reviewer C.