Received: 7 March 2022 | Accepted: 23 September 2022

DOI: 10.1111/2041-210X.14005

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Voice activity detection in eco-acoustic data enables privacy protection and is a proxy for human disturbance

Benjamin Cretois[1] | **Carolyn M. Rosten**[1] | **Sarab S. Sethi**[1,2,3]

[1]Norwegian Institute for Nature Research, Trondheim, Norway

[2]Department of Plant Sciences, University of Cambridge, Cambridge, UK

[3]Centre for Biodiversity and Environment Research, UCL, London, UK

Correspondence
Benjamin Cretois
Email: benjamin.cretois@nina.no; bencretois@gmail.com

Sarab S. Sethi
Email: sss70@cam.ac.uk

Funding information
Norges Forskningsråd, Grant/Award Number: 160022/F40

Handling Editor: Miguel Acevedo

Abstract

1. Eco-acoustic monitoring is increasingly being used to map biodiversity across large scales, yet little thought is given to the privacy concerns and potential scientific value of inadvertently recorded human speech. Automated speech detection is possible using voice activity detection (VAD) models, but it is not clear how well these perform in diverse natural soundscapes. In this study we present the first evaluation of VAD models for anonymization of eco-acoustic data and demonstrate how speech detection frequency can be used as one potential measure of human disturbance.
2. We first generated multiple synthetic datasets using different data preprocessing techniques to train and validate deep neural network models. We evaluated the performance of our custom models against existing state-of-the-art VAD models using playback experiments with speech samples from a man, woman and child. Finally, we collected long-term data from a Norwegian forest heavily used for hiking to evaluate the ability of the models to detect human speech and quantify a proxy for human disturbance in a real monitoring scenario.
3. In playback experiments, all models could detect human speech with high accuracy at distances where the speech was intelligible (up to 10 m). We showed that training models using location specific soundscapes in the data preprocessing step resulted in a slight improvement in model performance. Additionally, we found that the number of speech detections correlated with peak traffic hours (using bus timings) demonstrating how VAD can be used to derive a proxy for human disturbance with fine temporal resolution.
4. Anonymizing audio data effectively using VAD models will allow eco-acoustic monitoring to continue to deliver invaluable ecological insight at scale, while minimizing the risk of data misuse. Furthermore, using speech detections as a proxy for human disturbance opens new opportunities for eco-acoustic monitoring to shed light on nuanced human–wildlife interactions.

KEYWORDS
anonymization, bioacoustics, eco-acoustics, human disturbance, machine learning, privacy

1 | INTRODUCTION

Land-use change and global warming are impacting the natural world at an ever-increasing rate (Newbold et al., 2015; Root et al., 2003). By exploiting advances in sensor technology and data analysis techniques, scientists are now able to monitor and understand the resulting ecological changes both in detail and at unprecedented scales (Gijzen, 2013). Eco-acoustic monitoring is a particularly promising monitoring approach which offers high resolution data across a wide range of taxa over long time periods (Gibb et al., 2019; Pijanowski et al., 2011). However, while the field of eco-acoustic monitoring is expanding rapidly, surprisingly little attention has been paid to an almost universal issue: the impact, implications and potential value of inadvertently recorded human speech.

Eco-acoustic data are collected easily and inexpensively and can be analysed with increasingly sophisticated analytical techniques (Aide et al., 2013; Hill et al., 2018; Sethi, Ewers, et al., 2020; Sethi, Jones, et al., 2020; Sueur et al., 2008). The maturity of the technology has led to autonomous monitoring networks being deployed across vast landscapes, and even across full nations (Roe et al., 2021; Sethi et al., 2021; Sethi, Ewers, et al., 2020; Sethi, Jones, et al., 2020). While the primary focus is to derive ecological data such as species distributions or community assemblages, in some cases recording of human speech is inevitable. The presence of identifiable human speech raises serious ethical questions regarding data privacy and opens the door to potentially nefarious uses of eco-acoustic data; yet this has not been discussed in the literature. Manual filtering of data or obtaining prior consent is not appropriate nor practically feasible in most cases, and therefore an automated approach to anonymizing eco-acoustic data is required.

Acoustic data can be anonymised by blurring (e.g. using a source separation approach; Cohen-Hadria et al., 2019) or simply removing sections of audio containing identifiable speech. Solutions to the most challenging step, voice activity detection (VAD), have been explored in other fields of acoustic data analysis (Ramirez et al., 2007; Sohn et al., 1999). However, applications tend to be either in situations where the speaker is close to or speaking directly into a microphone (Pfau et al., 2001), or in urban environments with relatively low levels of homogenous background noise (Cohen-Hadria et al., 2019). Detecting speech in highly diverse and noisy passively collected eco-acoustic data is a more difficult task; especially since common components such as bird calls can overlap in frequency with speech and may trigger false positives (Hu & Cardoso, 2009). In-depth evaluation of different data processing techniques and adaptation of existing state-of-the-art VAD models is required to ensure anonymization can be done reliably given the uniquely challenging nature of eco-acoustic data.

Recorded human speech, while undesirable from a privacy standpoint, has the potential to serve as an invaluable proxy for human disturbance on an ecosystem (Gaynor et al., 2018). Eco-acoustic data have previously been used to estimate levels of human disturbance at a coarse soundscape level (e.g. Buxton et al., 2017, and CityNet from Fairbrass et al., 2019). Nonetheless, as opposed to existing approaches which group all anthropogenic acoustic activity into one class, voice activity detection could provide more precise information on exactly when and where humans were present (excluding sounds made by distant or passively operated equipment) as well as presenting the opportunity to estimate human population sizes and demographics (Bahari et al., 2014; Reynolds, 1995). Quantifying an aspect of human disturbance at this resolution and level of detail may shed light on subtle human–wildlife interactions that would otherwise be hidden when using coarser measures of disturbance.

In this study we present a novel approach to voice activity detection which is tailor-made for anonymization of eco-acoustic data. We trained convolutional neural networks (CNNs) on synthetic datasets comprised of human voices mixed with typical background sounds encountered in eco-acoustic data and evaluated performance with two distinct experiments. First, we performed playback experiments to measure our model's ability to detect different types of speech in a controlled manner and to test generalizability across landscapes. In the second experiment we collected longer-term data from an area regularly used for recreational hiking purposes to measure performance in a passive monitoring scenario and to explore the potential for VAD models to serve as a proxy for human disturbance. Both experiments were carried out in the Norwegian landscape. Through both experiments, we evaluated the effects of different data preprocessing approaches and compared performance to existing state-of-the-art VAD models; allowing us to provide recommendations on the best practices for safe and accurate automated voice detection in eco-acoustic data.

2 | METHODS

2.1 | Data collection

2.1.1 | Experiment 1: Audio data for evaluation using playback experiments

We selected two sites located in Børsa, Norway to collect training, validation and test audio data for playback experiments. Each site represented a different landscape type, namely a forest and a semi-natural grassland. At each site we deployed an Audiomoth version 1.1.1 device (Hill et al., 2018) that recorded audio data for five consecutive days during July 2021. The Audiomoth version 1.1.1 uses a micro-electro mechanical systems (MEMS) microphone that have a minimum, typical and maximum sensitivity of −21, −18 and −15 dBV/PA respectively with sensitivities given for a reference condition of 94 dB Sound Pressure Level at 1 kHz. Thus, a total of 10 days of audio files was collected (5 days for both forest and semi-natural grassland). We placed the audio recorders at least 100 m from the closest hiking trail to avoid unnaturally high levels of human noise pollution. Audio was recorded at a sampling frequency of 44.1 kHz continuously in chunks of 55 s and files were saved in the WAV format. For both forest and semi-natural grassland soundscape, 4 days (N = 5760 soundscape audio files) were used for training and 1 day ($N = 1440$ soundscape audio files) was retained for validation.

To test the VAD models, we recorded speech from three different speakers (a female, a male and a child) in a soundproof environment. Each person read the exact same sentence at a volume used during a normal conversation (Supplementary A) resulting in a recording of approximatively 30 seconds. We played the recorded speech using a JBL Xtreme 2 portable speaker, with the volume calibrated such that the male voice registered at 60 dB SPL (i.e. the volume of real speech).

Playbacks were performed during daytime (i.e. between 10 am and 19 pm) at distances of 1, 5, 10 and 20 m from an Audiomoth device in both the forest and semi-natural grassland locations. Samples of speech from each of the three speakers were played at each distance, both facing and from behind the recording device. In total, we collected 48 two-min recordings, each containing approximatively 30 s of speech and 1 min 30 s of ambient soundscapes. The recorded audio was divided into overlapping 3 s clips and each was manually labelled as 'speech' or 'no speech'. The 3 s segments were labelled as 'speech' if they contained any human speech, regardless of whether it was the start, middle or end of an utterance.

2.1.2 | Experiment 2: Evaluation in a passive monitoring scenario

To measure the performance of our model in a passive monitoring scenario we collected data from the Bymarka forest near Trondheim, Norway. This forest is particularly popular for weekend hikes and outdoors activities, so we expected significant human activity and speech. Because of data privacy regulations, Trondheim municipality provided a research permit stating that we can monitor and use audio data gathered in the Bymarka forest (permit reference 21 /40874, delivered by Trondheim municipality; Supplementary B). Two Audiomoth version 1.1.1 devices were deployed for 5 days each (from 30 September 2021 to 4 October 2021) and data were recorded in files of 55 seconds at a sampling frequency of 44.1 kHz. The first location (Forest 1) was located closer to a road (distance from the nearest road = 40 m) while the second location (Forest 2) was located further into the forest (distance from the nearest road = 140 m).

The acoustic dataset contained a total of 10 days (five per audio recorder) that were divided into a training, a validation and a test dataset. As traffic in the Bymarka forest is heightened during the weekends as opposed to weekdays, we retained the Saturday 2nd and Sunday 3rd of October 2021 as test dataset (4 days in total; $N = 5760$ soundscape audio files) as a way to both limit the amount of unwanted speech in the training dataset and to make sure speech was present on the test dataset. The training dataset was composed of five of the remaining days ($N = 7200$ soundscape audio files) while the validation dataset was composed of 1 day only ($N = 1440$ soundscape audio files).

It was not possible to manually label every instance of speech in the Bymarka dataset due to the sheer volume of data. However, we estimated the proportion of human speech in the dataset by randomly sampling 100 raw audio files of 55 s from the test dataset (i.e. Saturday 2nd and Sunday 3rd of October) and manually listening for any audible voices.

2.2 | Data preprocessing pipeline

Ecosystems have soundscapes which are far more diverse than the environments in which VAD models are typically deployed. Common sounds come from both biotic (e.g. species vocalizing or moving) and abiotic (e.g. rain, wind, motors) sources—each of which can vary from site to site and can confound sound event detection models. Existing human speech datasets used to train VAD models, however, do not capture this diversity of background sounds. Therefore, we combined several datasets to train a custom CNN-based VAD model on a large synthetic dataset which is representative of typical eco-acoustic data (Salamon & Bello, 2017).

To create the synthetic dataset, we split the raw soundscape recordings into nonoverlapping 3 s segments. Then, each segment was mixed with either both human speech and background noises (e.g. wind, birds, etc.), human speech only, background noises only or remained unmixed.

Human voices were randomly sampled from the LibriSpeech dataset (Librispeech, Panayotov et al., 2015). We chose LibriSpeech as the sole dataset for human speech as it contains about 360 hours of cleaned 16 kHz English speech with male and female voices in equal proportion. LibriSpeech is typically used to train models for voice activity detection or segmentation tasks in less noisy environments (Panayotov et al., 2015). Background noises included animal vocalizations (from BirdClef2017, Kahl et al., 2018), and environmental and anthropogenic sounds (from ESC50, Piczak, 2015; Supplementary C). We added both human speech and background noises to 5%, human speech only to 45%, and background noises only to 25% of the 3 s records. We did not add any human speech or background noises to 25% of the records (Figure 1a). The ratios used were decided based on preliminary explorations of model performance on the validation dataset using alternative ratios.

When only one sound type was added (speech or background noise) waveforms were scaled by a parameter, α, such that the amplitude of the added audio varied randomly in the range [−56.16, −8.3] dBFS (Equation 1, Supplementary D). To remain consistent with normal expected soundscape amplitudes, the range was chosen based on the minimum and maximum dBFS values of the recorded data. If background noise and human speech were both added to a 3 s audio file, a parameter β drawn from a uniform distribution in the range [0.1, 0.9] was additionally used (Equation 2).

$$\text{Mix} = \alpha\,(\text{human speech OR background noise}) + \text{soundscape}, \quad (1)$$

$$\text{Mix} = \alpha \times (\beta \times \text{human speech} + (1 - \beta) \times \text{background noise}) + \text{soundscape}, \quad (2)$$

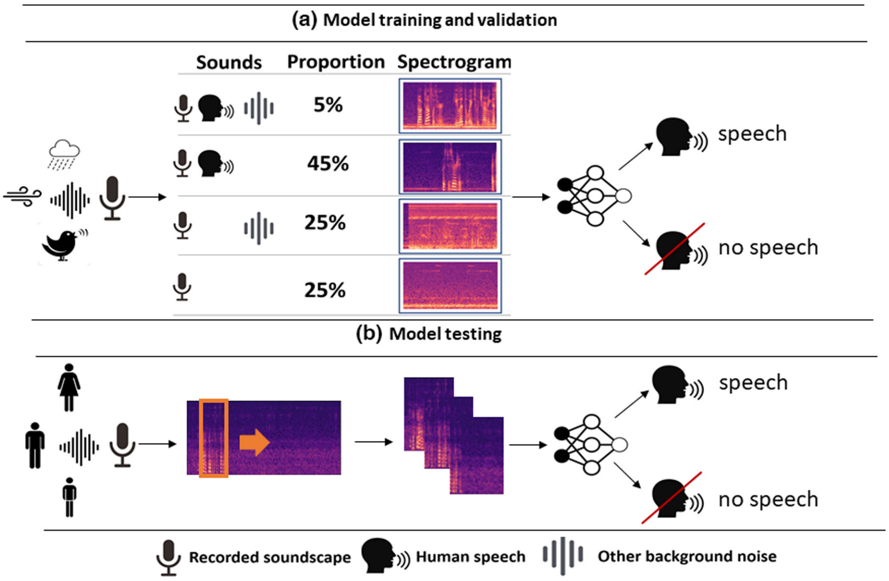

FIGURE 1 Schematic overview of our approach to automated human speech detection in eco-acoustic data. (a) Training and validation data were created using a data preprocessing approach, where existing datasets of human speech and typical background noises were overlaid with varying probabilities on baseline ecosystem soundscape data. A convolutional neural network was trained to detect human speech using the preprocessed data. (b) We performed playback experiments using recorded speech samples from a man, woman and child to measure the accuracy of our model. We also evaluated the performance of our model on longer passively recorded eco-acoustic data.

From the raw samples in the Librispeech, BirdClef2017 and ESC50 datasets, we randomly subsampled 3 s of audio, so we did not only capture the beginnings of the samples. To simulate partial captures, we also included a probability of random shifts in start time for both the human speech and background noises. Thus, speech or background noise clips were started anytime between [0, 2] seconds in the 3 s segment for a duration of at least 1 s. We also added a fade in and fade out effect of 500 ms for the human speech clips to simulate a human speaker moving towards or past the recording device. All the previously described steps, namely background noise mixing, sound to noise ratio modification and other sound effects such as fade in/fade out are referred to as 'data preprocessing'. Data preprocessing can artificially make the bioacoustics training set more diverse. Diversifying the dataset is crucial for the model to generalize better to soundscapes it has not been trained on, and this strategy has also been shown to be beneficial for model performance (Stowell, 2022). For simplicity, we refer to the entire data preprocessing pipeline as ecoVAD.

We tested the efficacy of different data preprocessing strategies by training and validating models on a dataset with a silent background (i.e. a control dataset) and datasets using different preprocessing (i.e. treatment datasets). More specifically, the training and validation datasets were composed of (a) a fully preprocessed dataset (ecoVAD), (b) a dataset based on natural soundscape background with no extra background noises mixed in (e.g. wind, rain, birds, etc.; Soundscape BG), (c) a dataset based on white-noise background with extra background noises added (White-noise BG + noise) and (d) the control, a dataset based on white-noise background with no extra background noises mixed in (White-noise BG). By substituting the natural soundscapes for white-noise in two of the cases (White-noise BG + noise, and White-noise BG) it was possible to test the effect of using location-specific soundscape data on the accuracy of the model. A white-noise background was used rather than a silent background to control for signal to noise ratio in the final processed sample.

To generate Soundscape BG we removed all background noises from the pipeline, the dataset thus containing 50% of segments with human speech and soundscape background with varying SNR (Equation 1) and 50% of segments with unmodified soundscape. White-noise BG + noise was generated by replacing the soundscape background by white-noise whose volume was fixed at −50 dBFS, the root mean square (RMS) volume of the raw recorded soundscapes. The rest of the pipeline was unchanged. Finally, White-noise BG was generated by removing all background noises and replacing the soundscape by white-noise as described above. We ensured that the same number of training and validation files was used for all processing types and that, where appropriate, each processing type included the same speech and background noise files.

We converted each preprocessed 3 s segment into a Mel spectrogram that was used as an input to the model. Acknowledging that our binary classification task did not require very high temporal resolution, we selected a Fast Fourier Transform (FFT) window size of 64 ms (1024 samples at 16 kHz) and an overlap of 50% (hop length of 512). We performed frequency compression using a Mel scale with 128 bands. Finally, we normalized the Mel spectrograms along each frequency bin, as this step has been shown to significantly improve classifier performance for audio classification tasks (Nagrani et al., 2017). These steps resulted in a Mel spectrogram of dimension 128 × 128 pixels that was used as an input for the convolutional neural network model (Figure 1a).

2.3 | CNN model architecture and training

All custom CNN models were based on the VGG11 architecture (Simonyan & Zisserman, 2015). We changed the number of input neurons so that the model accepted fixed-size 128 × 128 images containing a single colour channel, and the number of output neurons on the final fully connected layer was set to one. This way, the model outputs values close to 1 if speech has been detected or close to 0 if no speech has been detected. To reduce overfitting, we also added batch normalization after each convolutional layer and a dropout of 0.5 for the fully connected layers (for a visual description of the model see Figure S1). As the classification task is binary (speech vs. no speech) we used a binary cross entropy loss function. More details concerning the software used can be found in Supplementary E.

Because the experiment 1 dataset contained two different landscapes (forest and semi-natural grassland) we trained one model for each data preprocessing type (i.e. ecoVAD, White-noise BG + noise, Soundscape BG and White-noise BG) and for each landscape type. We also trained models using data from both landscapes put together (forest + semi-natural grassland). This resulted in a total of 12 models for the playback dataset.

For experiment 2, one model per data preprocessing type (i.e. ecoVAD, White-noise BG + noise, Soundscape BG and White-noise BG) was trained. This resulted in a total of four models. It should also be noted that the raw soundscape recordings used for experiment 2 may have contained some speech, which, after being preprocessed by our pipeline, could end up being labelled as 'no speech'. Nevertheless, as opposed to the test set, the training set was collected during weekdays, and we did not expect a low proportion of human speech to make a significant difference in model training.

All models were trained using a learning rate of 0.1, 0.01 and 0.001 with a decay of 0.1 every 20 epochs. An early stopping strategy based on the validation loss was used to stop the training of the model if it was complete before the maximum of 200 epochs. Performance of the models was evaluated on the validation dataset, and we used the area under the receiver operating characteristic curve (ROC AUC, herein AUC) to select the best performing learning rates per data preprocessing type and landscape type. The performances of the best models were then evaluated on the test datasets.

2.4 | CNN model testing

Model predictions on the test data for both experiments 1 and 2 were computed using a sliding window with an overlap of 2 s, resulting in one prediction per second of audio (Figure 1b).

Since we had fully labelled data for experiment 1, we were able to measure AUC for each model. However, similar ground truth labels were unavailable for experiment 2 as the length of the recording period made manual labelling infeasible. Therefore, to estimate model performance for each processing type, we randomly selected 100 detections per Audiomoth that we manually listened to confirm whether the detections were correct (i.e. contained any human speech). Comparing the numbers of true positives and false negatives yielded the precision for each model.

In addition to evaluating the potential of our approach for anonymizing soundscape records, we also showed the potential of such a model to provide a proxy for human disturbance upon the landscape by counting the number of detections the model made per hour. The number of detections was normalized for each site using min-max scaling to produce a proxy measure of relative human disturbance across the test period. To estimate whether the number of detections made through the day and their timing were coherent we compared the number of detections of the models with the arrival and departure of buses at the nearest bus stop from the Audiomoth (distance to nearest bus stop = 135, 235 m for forest 1 and forest 2 respectively). The bus timetable was obtained from the bus company website (https://www.atb.no/en/).

2.5 | Comparison with other VAD models

We assessed the performance of our pipeline by comparing the trained model with two existing state-of-the-art VAD models. Based on their widespread usage and differing approaches, we used pyannote v1.1.1 (Bredin et al., 2020), a CNN based on the PyanNet architecture, and Google WebRTC VAD v2.0.10 (Google WebRTC, n.d.), a Gaussian mixture model-based approach. In addition to the ecoVAD pipeline which can be fully customized for specific datasets (*VAD_algorithms/ecoVAD/train_model.py*), we provide wrappers around pyannote (*VAD_algorithms/pyannote/pyannote_predict.py*) and WebRTC VAD (*VAD_algorithms/webrtcvad/webrtcvad_predict.py*) in our GitHub repository. Performance was compared using the F1 score on the experiment 1 dataset across distances. The F1 score is a metric commonly used in machine learning to compare different classification models as it combined the precision and recall of a classifier into a single metric by taking their harmonic mean. We also assessed the precision of pyannote and WebRTC VAD on the dataset collected for experiment 2 following the protocol described in Section 2.5.

3 | RESULTS

3.1 | Experiment 1: Performance in controlled playback experiments

The model trained using the ecoVAD pipeline was able to classify human speech with a high degree of confidence up to 10 m for samples from a man, woman and child (classification confidence >0.8 ±0.170 SD for all speakers at 10 m; Figure 2a). At 20 m the classification confidence for speech decreased for all speakers (classification confidence = 0.727, 0.700, 0.707 ± 0.183, 0.210, 0.181 SD for man, woman and child respectively) and the speech score of the lower quartile overlapped with the scores for the nonspeech samples. However, at 20 m the speech was barely audible in the audio

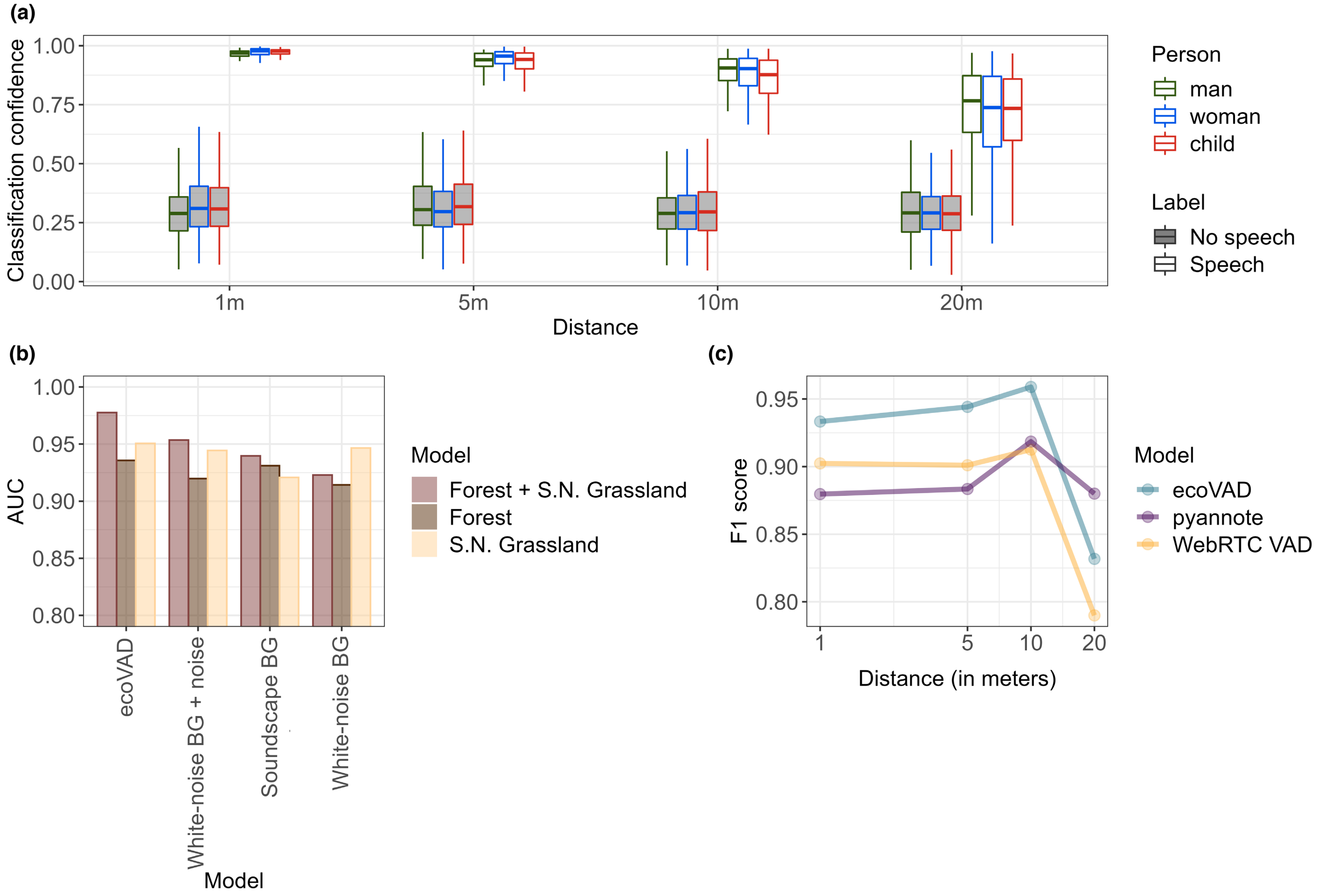

FIGURE 2 Voice activity detection (VAD) models performed well on eco-acoustic data in controlled playback experiments. (a) For the model trained using the ecoVAD pipeline, classification confidence is plotted against distance for samples of speech from man, woman and child. The model was able to successfully discriminate between 3 s audio segments with and without speech at distances of up to 20m (at which point the speech was unintelligible). Performance was consistent across samples spoken by a man, woman and child. (b) Accuracy of the models trained using different data preprocessing approaches and different soundscape types was consistently high (>0.9 AUC in all cases). A slight increase in performance was seen for models trained with soundscape data from both forest and semi-natural grasslands (S.N. Grassland) and when both noise and soundscape preprocessing techniques were used (ecoVAD). (c) The model trained using the ecoVAD pipeline outperformed two existing state-of-the art VAD models (WebRTC VAD and pyannote) when their performance was evaluated on the same playback experiments.

samples, and the sentences being spoken were not intelligible. The classification confidence for periods of audio with no speech in them remained low at all distances with a mean classification confidence of 0.30 ± 0.125 SD.

All models employing some form of data preprocessing (White-noise BG + noise, Soundscape BG, and ecoVAD) that were trained on both forest and semi-natural grassland reached higher AUCs than the model that was trained with no data preprocessing (Figure 2b). We did not notice dramatic differences in performance when testing a model on the landscape it has not been trained on. For instance, the model trained on the forest only data and the model trained on the semi-natural grassland only data had an AUC of 0.97 and 0.98 respectively when tested on forest data and an AUC of 0.91 and 0.93 respectively when tested on semi-natural grassland (Figure S2). While the model resulting from the ecoVAD pipeline was trained on a limited amount of data (i.e. only 4 days of soundscapes), it outperformed state of the art models such as pyannote and WebRTC VAD on the playback dataset (average F1 score = 0.917, 0.890 and 0.876 for ecoVAD, pyannote and WebRTC VAD respectively; Figure 2c). We noticed that the performance of all models decreased at 20 m, particularly for WebRTC VAD (F1 scores at 20 m = 0.832, 0.88 and 0.790 for ecoVAD, pyannote and WebRTC VAD respectively).

3.2 | Experiment 2: Using voice detection frequency as a proxy for human disturbance

In the passive monitoring dataset collected from Bymarka, Norway, again, data preprocessing improved the precision of models. By listening to a random subset of the audio, we found speech was present in approximately 5% of the test dataset. Therefore, a null model which selects audio at random would have a precision of 0.05 in this experiment. At both locations, the models trained without any data preprocessing or with only background noises added failed to outperform the null model (precision < 0.05, Figure 3a). The addition of soundscapes, however, resulted in a marked improvement in precision (0.25 in Forest 1, 0.18 in Forest 2). The addition of both background noises and soundscapes used in the ecoVAD pipeline resulted in further increased precision of the model in Forest 2 (precision = 0.32), but in Forest 1 resulted in a slight decrease in precision compared to the model trained without extra background noises (precision = 0.14).

In Forest 2, located 140 m from the nearest road, the model trained using the ecoVAD pipeline outperformed both pyannote and WebRTC VAD (increased precision of 0.05 compared to pyannote and 0.31 compared to WebRTC VAD, Figure 3a). However, pyannote outperformed both all models trained using the ecoVAD pipeline (regardless of the preprocessing approach used) and WebRTC VAD when evaluated on data from Forest 1 which was 40 m from the nearest road.

We found a strong link between bus timings and the number of hourly detections using both pyannote and the model trained using ecoVAD pipeline for both Forest 1 and 2. For WebRTC VAD the link between bus timing and number of hourly detections was clear for Forest 1 but less so for Forest 2, most likely due to the low precision of the model at this location. Between the hours of 06:00 and 18:00 inclusive (approximate daylight hours) there were more speech detections from both model trained using ecoVAD pipeline and pyannote than during the night-time hours (90%, 85% for ecoVAD and 91%, 97% for pyannote on Forest 1 and Forest 2 respectively). Both models registered peaks in normalized detections/hour between 06:00 and 12:00 on both 2 and 3 October at both locations. This is both supported by a higher frequency of bus arrivals in the morning and agrees with the intuition that hiking activity is likely to be higher in the morning than in the late afternoon and evening hours. While most of WebRTC VAD detections were located between 6:00 and 18:00 for Forest 1 (86%) this proportion decreased for Forest 2 (63%).

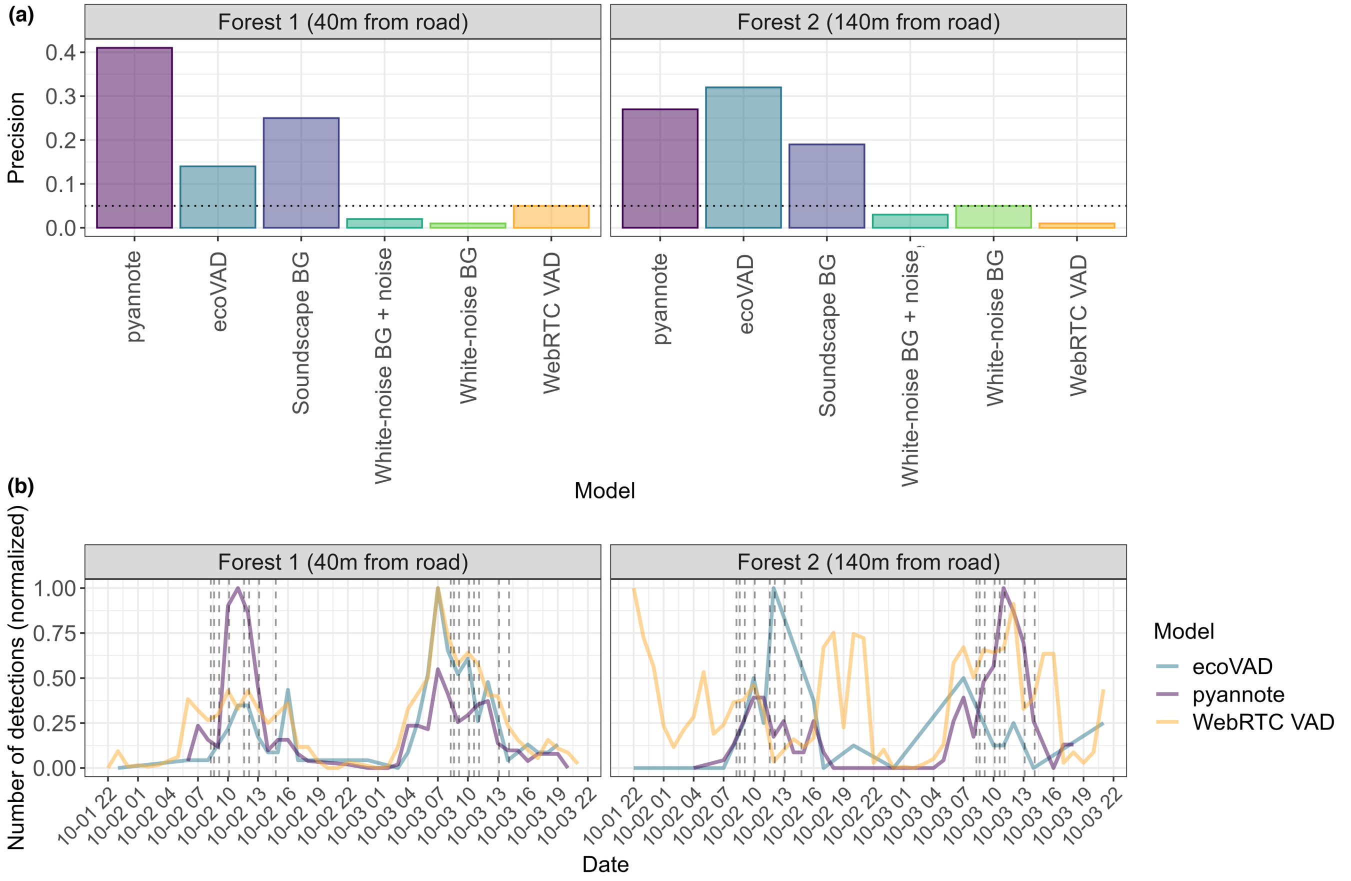

FIGURE 3 Frequency of voice activity detections serves as an accurate proxy of human disturbance on a fine temporal resolution. (a) One hundred automated voice detections from 2 days (2–3 October 2021) of eco-acoustic monitoring were manually labelled to derive the precision of five models at two sites in Bymarka, Norway. The addition of soundscapes had a larger impact upon model precision than the addition of background noises. Our model, trained using the ecoVAD pipeline, achieved state-of-the-art performance, outperforming WebRTC VAD at both locations and pyannote in Forest 2. WebRTC VAD and the models trained on White-noise BG + noise and White-noise BG performed worse than a null model (displayed as the dotted line) (b) Min-max normalized number of hourly detections across the same 2-day period for the model trained using ecoVAD pipeline (purple), pyannote (green) and WebRTC VAD (orange) are depicted alongside black dashed lines, which show bus arrival and departure times. Dates are formatted as month–day–hour. There are clear peaks in both voice detections and bus timings during the morning hours on both days, indicating that voice detections may be used as a fine scale proxy of human disturbance.

4 | DISCUSSION

In this study we showed that VAD models can be successfully used for human speech detection in eco-acoustic datasets. These detections can be used to silence sections of speech within the audio (see the *anonymise_data.py* example we provide) to ensure anonymity but can also be used for quantifying an aspect of human disturbance within an ecosystem on a fine spatial and temporal resolution. However, our results suggest that not all available models should be considered for the task of anonymization and human disturbance quantification of eco-acoustic data. We demonstrated that training a model using our ecoVAD pipeline, namely mixing publicly available speech datasets with soundscape recordings and common noise sources (e.g. wind, birdsongs) resulted in speech detection models which outperformed the state-of-the-art. A fundamental advantage of using the ecoVAD pipeline rather than an off-the-shelf model is that it makes it possible to create a model that is trained on a dataset of our choice, and thus results in improved accuracy (we provide a script train_ecovad.py in the ecoVAD repository to train a model on any custom dataset). As recreational users commonly take the bus to get to the Bymarka forest, we compared the frequency of speech detections with bus timings and found that the peaks in number of voice detections coincided with peak hours of bus activity. This suggests that voice activity detection models can be used as a proxy for human disturbance of an ecosystem on a fine temporal resolution.

The two applications of VAD algorithms to eco-acoustic data that we presented, data anonymization and human disturbance quantification, are quite different in nature. In this study, we used the same models for both tasks, and therefore selected a detection threshold that was a good trade of between recall (i.e. minimizing false negatives) and precision (i.e. minimizing false positives). Other use cases will call for different balances to be struck; if data are highly sensitive a lower detection threshold should be preferred, whereas, if the goal is to use VAD models for quantifying human disturbance higher thresholds may yield less noisy results.

Data preprocessing was used in our study when training our CNN models. Our playback experiments were conducted in the relatively silent soundscapes of Børsa, Norway, and, intuitively, we found that the data preprocessing step resulted only in a slight improvement of model accuracy in this setting. However, we found that data preprocessing vastly improved our model's precision for the passive monitoring data collected from the Bymarka forest. This was most likely due to the more diverse set of confounding sounds such as cars, motorcycle, footsteps, wind and animals contained within the longer recordings. Our results agree with prior work exploring the benefits or adding variety to the datasets in acoustic machine learning tasks, and suggest that data preprocessing may be even more important when working in more acoustically complex environments (e.g. the tropics; Stowell, 2022) where confounding sounds can be even more varied. Conversely, in less noisy environments off-the-shelf VAD models may be performant enough, avoiding the need to train deployment-specific models altogether.

Throughout our study we compared the performance of the models resulting from the ecoVAD pipeline with two existing state-of-the-art models commonly used for voice activity detection; pyannote and WebRTC VAD. In the playback experiments we found that our model consistently outperformed both WebRTC VAD and pyannote (at all distances except at 20 m, at which point the speech was unintelligible to a human listener). However, in the passive monitoring scenario, while our model had higher precision than WebRTC VAD in both locations, pyannote outperformed our model at the location nearer to the road. Analysing the sources of false detections hinted that pyannote was less likely than our model to falsely identify anthropogenic sounds as speech, but more likely to be confused by natural sounds such as wind or branches falling (Figure S3). Even though the ecoVAD pipeline used anthropogenic noise sources to train our model, the data preprocessing was heavily weighted towards biotic sources, while the training data used for pyannote (e.g. VoxCeleb; Bredin et al., 2020) will have contained a greater diversity of anthropogenic background sounds. It should, however, be noted that with models exclusively trained with anthropogenic background sounds (CityNet for instance; Fairbrass et al., 2019) outputs are likely to include a large number of false positives (Figure S4). WebRTC VAD, a classical machine learning algorithm, has been especially trained for voice active detection in real-time (Google WebRTC, n.d.) and trades accuracy for speed. As shown in our study, WebRTC VAD may not be appropriate for post-hoc analyses on large datasets where speed of computation is less important, as it returns a large number of false detections. There is unlikely to be a one-size-fits-all approach, however, and further work may either employ ensemble approaches or find a better balance in the data preprocessing process to enable more robust voice detection across a wide variety of soundscapes.

We found that the frequency of voice detections per hour was linked closely to bus activity at a popular hiking spot in Bymarka, Norway. However, it could be argued that the model could also trigger on bird songs and not only on human voices as both have a spike of activity during the day. Listening to a sample of false detections revealed that our model's false positives are triggered by animal sounds in only 5% of the cases (Figure S3). The vast majority of our model's false detections were triggered by sounds that we could not hear or identify (Figure S3). It could simply be slight wind gusts that were undetectable to our ears, or other discrete background noises. In the case where one suspects wind to be a major issue, it is possible to implement a denoising strategy to remove some of it (Juodakis & Marsland, 2021). On the other hand, if the false detections are from discrete background noises, a possible remedy would be to, before training a model, assess the full spectrum of background noises of an ecosystem and add background noises in the pipeline relative to this assessment. The model would likely be more robust and yield fewer false positives.

Our results demonstrate that speech detections can be used as both a direct measure of anthropogenic noise pollution and an indirect proxy of human disturbance on an environment—both of which impact upon biodiversity (Boivin et al., 2016; Sordello et al., 2020). Furthermore, similar models could be developed to detect other anthropogenic sounds sources (e.g. vehicles, building activity) building out a fuller picture of human disturbance. Using eco-acoustic data to simultaneously track biodiversity alongside proxies of human disturbance presents a unique opportunity to gain a nuanced understanding of human-wildlife interactions on fine temporal scales; with the potential to inform more sustainable conservation programs and land management practices.

AUTHOR CONTRIBUTIONS

All authors contributed to the conceptualization of the manuscript. Carolyn M. Rosten acquired funding and provided technical support. Benjamin Cretois and Sarab S. Sethi analysed the data and wrote the original draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version.

ACKNOWLEDGEMENTS

We thank Wouter, Raghnild and Marius Koch who volunteered to record the samples for the playback experiments. We thank Frøde Fossoy who took part in the initial discussions that resulted in this work. We also thank Francesco Frassinelli who gave support for running the pipeline on High Performance Clusters. Trondheim commune gave permission for recording of audio data in the Bymarka forest (permit reference 21/40874). This work was supported by the Norwegian Research Council (grant number 160022/F40). S.S.S. was supported by a Herchel Smith Fellowship from the School of Biological Sciences at the University of Cambridge.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.14005.

DATA AVAILABILITY STATEMENT

Code for ecoVAD is archived in Zenodo (Cretois et al., 2022), and kept up to date on https://github.com/NINAnor/ecoVAD. Some data have been shared to facilitate reproducibility, but due to privacy concerns full audio datasets cannot be shared. Both data and model weights are available on OSF and can be found at https://osf.io/f4mt5/. We facilitated easy use and modification of our pipeline by providing scripts for end-to-end training and evaluation of models and configuration files that users can modify to train a model adapted to their own needs.

ORCID

Benjamin Cretois https://orcid.org/0000-0001-8668-3321
Carolyn M. Rosten https://orcid.org/0000-0002-1117-444X
Sarab S. Sethi https://orcid.org/0000-0002-5939-0432

REFERENCES

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. *PeerJ*, *1*, e103. https://doi.org/10.7717/peerj.103

Bahari, M. H., McLaren, M., Van Hamme, H., & van Leeuwen, D. A. (2014). Speaker age estimation using i-vectors. *Engineering Applications of Artificial Intelligence*, *34*, 99–108. https://doi.org/10.1016/j.engappai.2014.05.003

Boivin, N. L., Zeder, M. A., Fuller, D. Q., Crowther, A., Larson, G., Erlandson, J. M., Denham, T., & Petraglia, M. D. (2016). Ecological consequences of human niche construction: Examining long-term anthropogenic shaping of global species distributions. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(23), 6388–6396. https://doi.org/10.1073/pnas.1525200113

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P. (2020). Pyannote. Audio: Neural building blocks for speaker diarization. *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7124–7128). https://doi.org/10.1109/ICASSP40776.2020.9052974

Buxton, R. T., McKenna, M. F., Mennitt, D., Fristrup, K., Crooks, K., Angeloni, L., & Wittemyer, G. (2017). Noise pollution is pervasive in U.S. protected areas. *Science*, *356*(6337), 531–533. https://doi.org/10.1126/science.aah4783

Cohen-Hadria, A., Cartwright, M., McFee, B., & Bello, J. P. (2019). Voice anonymization in urban sound recordings. *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). https://doi.org/10.1109/MLSP.2019.8918913

Cretois, B., Rosten, M. C., & Sethi, S. S. (2022). ecoVAD: An end to end pipeline for training and using VAD models in soundscape analysis. *Zenodo*, https://doi.org/10.5281/zenodo.7137250

Fairbrass, A. J., Firman, M., Williams, C., Brostow, G. J., Titheridge, H., & Jones, K. E. (2019). CityNet—Deep learning tools for urban ecoacoustic assessment. *Methods in Ecology and Evolution*, *10*(2), 186–197. https://doi.org/10.1111/2041-210X.13114

Gaynor, K. M., Hojnowski, C. E., Carter, N. H., & Brashares, J. S. (2018). The influence of human disturbance on wildlife nocturnality. *Science*, *360*(6394), 1232–1235. https://doi.org/10.1126/science.aar7121

Gibb, R., Browning, E., Glover-Kapfer, P., & Jones, K. E. (2019). Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, *10*(2), 169–185. https://doi.org/10.1111/2041-210X.13101

Gijzen, H. (2013). Big data for a sustainable future. *Nature*, *502*(7469), 38. https://doi.org/10.1038/502038d

Google WebRTC. (n.d.). https://webrtc.org/

Hill, A. P., Prince, P., Covarrubias, E. P., Doncaster, C. P., Snaddon, J. L., & Rogers, A. (2018). AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. *Methods in Ecology and Evolution*, *9*(5), 1199–1211. https://doi.org/10.1111/2041-210X.12955

Hu, Y., & Cardoso, G. C. (2009). Are bird species that vocalize at higher frequencies preadapted to inhabit noisy urban areas? *Behavioral Ecology*, *20*(6), 1268–1273. https://doi.org/10.1093/beheco/arp131

Juodakis, J., & Marsland, S. (2021). Wind-robust sound event detection and denoising for bioacoustics. *ArXiv:2110.05632 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/2110.05632

Kahl, S., Wilhelm-Stein, T., Klinck, H., Kowerko, D., & Eibl, M. (2018). Recognizing birds from sound—The 2018 BirdCLEF baseline system. *ArXiv:1804.07177 [Cs]*. http://arxiv.org/abs/1804.07177

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. *Interspeech*, *2017*, 2616–2620. https://doi.org/10.21437/Interspeech.2017-950

Newbold, T., Hudson, L. N., Hill, S. L. L., Contu, S., Lysenko, I., Senior, R. A., Börger, L., Bennett, D. J., Choimes, A., Collen, B., Day, J., De Palma, A., Díaz, S., Echeverria-Londoño, S., Edgar, M. J., Feldman, A., Garon, M., Harrison, M. L. K., Alhusseini, T., … Purvis, A. (2015).

Global effects of land use on local terrestrial biodiversity. *Nature*, *520*(7545), Article 7545. https://doi.org/10.1038/nature14324

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5206–5210). https://doi.org/10.1109/ICASSP.2015.7178964

Pfau, T., Ellis, D. P. W., & Stolcke, A. (2001). Multispeaker speech activity detection for the ICSI meeting recorder. *IEEE workshop on automatic speech recognition and understanding, 2001. ASRU'01* (pp. 107–110). https://doi.org/10.1109/ASRU.2001.1034599

Piczak, K. J. (2015, October). ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1015–1018). https://doi.org/10.1145/2733373.2806390

Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., Farina, A., Krause, B. L., Napoletano, B. M., Gage, S. H., & Pieretti, N. (2011). Soundscape ecology: The science of sound in the landscape. *Bioscience*, *61*(3), 203–216. https://doi.org/10.1525/bio.2011.61.3.6

Ramirez, J., Górriz, J. M., & Segura, J. C. (2007). Voice activity detection. Fundamentals and speech recognition system robustness. *Robust Speech Recognition and Understanding*, *6*(9), 1–22.

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, *17*(1), 91–108. https://doi.org/10.1016/0167-6393(95)00009-D

Roe, P., Eichinski, P., Fuller, R. A., McDonald, P. G., Schwarzkopf, L., Towsey, M., Truskinger, A., Tucker, D., & Watson, D. M. (2021). The Australian acoustic observatory. *Methods in Ecology and Evolution*, *12*(10), 1802–1808. https://doi.org/10.1111/2041-210X.13660

Root, T. L., Price, J. T., Hall, K. R., Schneider, S. H., Rosenzweig, C., & Pounds, J. A. (2003). Fingerprints of global warming on wild animals and plants. *Nature*, *421*(6918), 57–60. https://doi.org/10.1038/nature01333

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, *24*(3), 279–283. https://doi.org/10.1109/LSP.2017.2657381

Sethi, S. S., Ewers, R. M., Jones, N. S., Signorelli, A., Picinali, L., & Orme, C. D. L. (2020). SAFE Acoustics: An open-source, real-time eco-acoustic monitoring network in the tropical rainforests of Borneo. *Methods in Ecology and Evolution*, *11*(10), 1182–1185. https://doi.org/10.1111/2041-210X.13438

Sethi, S. S., Fossøy, F., Cretois, B., & Rosten, C. M. (2021). Management relevant applications of acoustic monitoring for Norwegian nature – The Sound of Norway. In *31*. Norsk institutt for naturforskning (NINA). https://brage.nina.no/nina-xmlui/handle/11250/2832294

Sethi, S. S., Jones, N. S., Fulcher, B. D., Picinali, L., Clink, D. J., Klinck, H., Orme, C. D. L., Wrege, P. H., & Ewers, R. M. (2020). Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(29), 17049–17055. https://doi.org/10.1073/pnas.2004702117

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ArXiv:1409.1556 [Cs]*. http://arxiv.org/abs/1409.1556

Sohn, J., Kim, N. S., & Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, *6*(1), 1–3. https://doi.org/10.1109/97.736233

Sordello, R., Ratel, O., Flamerie De Lachapelle, F., Leger, C., Dambry, A., & Vanpeene, S. (2020). Evidence of the impact of noise pollution on biodiversity: A systematic map. *Environmental Evidence*, *9*(1), 20. https://doi.org/10.1186/s13750-020-00202-y

Stowell, D. (2022). Computational bioacoustics with deep learning: A review and roadmap. *PeerJ*, *10*, e13152. https://doi.org/10.7717/peerj.13152

Sueur, J., Pavoine, S., Hamerlynck, O., & Duvail, S. (2008). Rapid acoustic survey for biodiversity appraisal. *PLoS ONE*, *3*(12), e4065. https://doi.org/10.1371/journal.pone.0004065

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Cretois, B., Rosten, C. M., & Sethi, S. S. (2022). Voice activity detection in eco-acoustic data enables privacy protection and is a proxy for human disturbance. *Methods in Ecology and Evolution*, *00*, 1–10. https://doi.org/10.1111/2041-210X.14005