TRANSFER LEARNING FROM YOUTUBE SOUNDTRACKS TO TAG ARCTIC ECOACOUSTIC RECORDINGS

Enis Berk Çoban[1], *Dara Pir*[2], *Richard So*[3], *Michael I Mandel*[4]

[1] The Graduate Center, CUNY, New York, NY
[2] Guttman Community College, CUNY, New York, NY
[3] Staten Island Technical High School, New York, NY
[4] Brooklyn College, CUNY, New York, NY

ABSTRACT

Sound provides a valuable tool for long-term monitoring of sensitive animal habitats at a spatial scale larger than camera traps or field observations, while also providing more details than satellite imagery. Currently, the ability to collect such recordings outstrips the ability to analyze them manually, necessitating the development of automatic analysis methods. While several datasets and models of large corpora of video soundtracks have recently been released, it is not clear to what extent these models will generalize to environmental recordings and the scientific questions of interest in analyzing them. This paper investigates this generalization in several ways and finds that models themselves display limited performance, however, their intermediate representations can be used to train successful models on small sets of labeled data.

Index Terms— Ecoacoustics, soundscape analysis, transfer learning

1. INTRODUCTION

Arctic-boreal forests in Alaska and Northern Canada are vast, remote, and relatively undisturbed, but they have been warming at a rate two to three times the global average [1]. At the same time, human development for resource extraction continues to encroach and intensify in these same areas. The inaccessibility of these regions presents challenges for the study of their ecosystems over their full spatial and temporal extents. Traditional *in situ* studies provide infrequent measurements at small spatial scales. Remote sensing using, e.g., satellite imaging, cannot capture information about wildlife behavior or phenology (the timing of life-cycle events). Autonomous audio recording networks, however, avoid both of these problems and are able to provide long-term observations over a large spatial extent.

These recordings can be used in the new field of soundscape ecology [2], which has used similar recordings in other ecosystems for biodiversity assessment [3], detecting threatened and invasive species [4], measuring levels of habitat destruction, fragmentation, and chemical pollution [5], and determining the abundance of bird species [6]. It explores the collection of biological, geophysical and anthropogenic sounds and the methods of processing their temporal, spatial and spectral characteristics to understand their association with ecological processes.

Most soundscape ecology studies, however, are still carried out using manual annotations collected from expert listeners reviewing these recordings at close to real-time speeds [7], limiting their ability to scale. For this reason, researchers are looking for ways to take advantage of machine learning algorithms for automating natural sound processing [8]. While it is still time-consuming and expensive to manually label training data for these methods, transfer learning [9] can reduce the amount of labeled data needed. Transfer learning involves the use of a model pre-trained on a large mismatched dataset to generate input features for another machine learning system trained on the target dataset. One popular pre-trained model for audio is VGGish [10].

This paper investigates whether the VGGish model and the related Audio Set dataset [11], both based on soundtracks of YouTube videos, can be effectively applied to the analysis of ecoacoustic soundscape recordings collected in the summer of 2016 from 20 sites along the Colville River and its delta in the North Slope of Alaska [12]. This area provides transportation to the oil and gas extraction activities in the Prudhoe Bay Oil Field. The village of Nuiqsut is situated on the Colville and is home to a Native Alaskan community that harvests caribou for their subsistence. Some harvesters have reported changes in caribou distribution and yields because of aircraft traffic in the area [13]. Thus, we analyze the presence of aircraft and other human-generated sounds in these recordings.

We also analyze the presence of bird sounds in these recordings. Every year millions of long-distance migratory birds travel from their overwintering grounds throughout North America to the North Slope of Alaska. There is emerging evidence that Arctic-breeding birds may be vulnerable to the fact that the timing of spring snow and ice cover melt is becoming increasingly heterogeneous in both time and space [14, 15]. Thus, the automatic identification of both songbirds and waterfowl in these recordings is of interest as well in identifying the timing of their spring and autumn migrations. Preliminary studies [16] have shown success in this endeavor using supervised and unsupervised approaches to classifying audio texture features [17]. The current study extends this to use deep learning models in the context of transfer learning.

Thus, in this study, we measure the ability of several different acoustic classifiers to recognize in these recordings sounds of eight different classes, including birdsong, waterfowl sounds, and aircraft noise. These acoustic classifiers include the bulbul model [18], an attention-based Audio Set classifier [19] with a manual mapping of tags to our classes, the same classifier with a learned mapping of tags, and a collection of standard supervised machine learning algorithms predicting our classes directly from the VGGish embeddings. A summary of these three systems is shown in Figure 1.

2. RELATION TO PRIOR WORK

The development of machine learning tools for applications in ecological studies using sound recordings is an interdisciplinary field

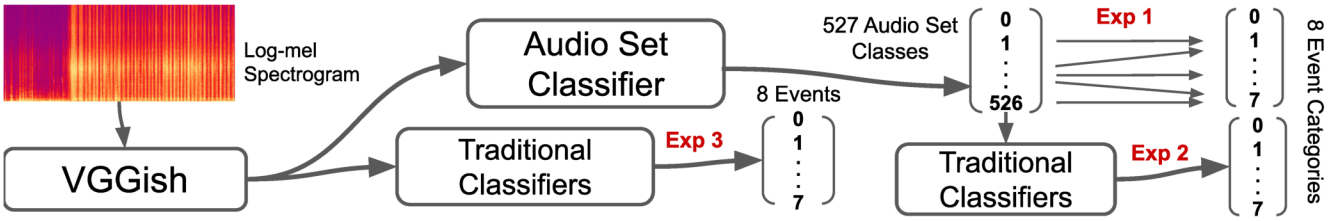

Fig. 1. System diagram of our three experiments

of research. Along with soundscape ecology, another term coined to bring such studies under one umbrella is ecoacoustics [20]. Ecoacoustics studies have also contributed to the automated calculation of indices such as the Acoustic Diversity Index and Bioacoustic Index [21]. Such studies have enabled the monitoring of animal biodiversity in tropical forests [22]. Classification of animal species is another common machine learning application using ecological sound data, including identifying species of birds [23, 24] and frogs [25], and detecting koalas [26]. Measuring the negative human impact on ecosystems in the context of global warming with ecological sound data is an active research area as well [27].

3. TECHNICAL DESCRIPTION OF SYSTEM

We use a pre-trained VGGish model [10] to generate audio embedding vectors from log mel spectrograms of 1-second sound excerpts. VGGish's architecture is similar to the vision system VGGNet [28]; VGGNet is a deep convolutional network originally developed for object recognition. VGGish's only difference is that it has 3087 sigmoid units in its output layer. While the original VGGNet has 144M weights and 20B multiplies, the audio variant uses 62M weights and 2.4B multiplies. VGGish is trained on the Youtube-100M dataset [29]. We use the model to predict 128-dimensional embedding vectors from non-overlapping 960 ms segments of audio. Each segment is processed with a short-time Fourier transformation with 25 ms windows calculated every 10 ms. The generated spectrograms are converted to 64 mel-spaced frequency bins and the magnitude of each bin is log transformed. At the end of this pre-processing, the log-mel spectrogram patches are of size 96×64.

As part of the same project, Google released a large labeled dataset of YouTube video soundtracks called Audio Set which is a collection of hierarchically organized sound events [11]. Audio Set covers many everyday sounds from human and animal activities to natural and environmental sounds [11]. This dataset is composed of human-labeled 10-second audio clips extracted from YouTube videos. The total size of the Audio Set dataset is over 1 million excerpts ($\sim$ 112 days), but it is still only 0.05% of the Youtube-100M dataset. Hershey et al. [10] find that predicting Audio Set labels from VGGish embeddings is more accurate than predicting them directly from spectrograms. Kong et al. [19] apply attention-based neural networks to VGGish embedding vectors for Audio Set classification and further improve performance. During our experiments, this was the best result in the literature, so we used this model to make Audio Set Ontology-based classifications on our data set.

Detection of bird sound events is of prime importance to our research. We assess the detection performance of the VGGish system by comparing its prediction patterns with those of the state-of-the-art bulbul system [18]. The bulbul system achieved the highest score in the 2017 Bird Audio Detection challenge [30]. In this paper, we use a modified version of bulbul, which was employed as the competitive baseline for the most recent Bird Audio Detection challenge in DCASE 2018. Bulbul generates predictions of the presence or absence of bird sound of any kind.

4. EXPERIMENTS

4.1. Data

Our experiments use ecological sounds collected by Taylor Stinchcomb [12] along the Colville River in Alaska. This dataset is composed of recordings made continuously in 20 different locations over a period of 3 months (June, July, and August of 2016) with total duration 837 days. Stinchcomb [12] used these recordings to manually investigate where and to what extent aircraft disturbance poses a threat to caribou harvest practices in Nuiqsut.

To assess the performance of the Audio Set classifiers on these recordings, three of the authors applied eight event categories to 1300 10-second clips sampled uniformly at random from them. We found the 10-second audio duration, which is used by the attention-based Audio Set predictor [19], to also be suitable for our annotation purposes. The eight categories are: wind, running water, rain, cable noise, songbird, waterbird, insect, and aircraft. The total number of positive instances of each label is shown in Table 1. We preferred high-level categories rather than specific species of animals to decrease the cost of manual labeling. Cable noise was caused by cables from the recorders banging against other pieces of equipment when blown by the wind.

Annotations were made using a simple GUI interface implemented in a Jupyter notebook showing the eight description categories. The annotators were allowed to listen to a clip as many times as needed and were asked to select all detected events. Our choice of tags was established by listening to random clips and enumerating all of the sounds present. We also provided a write-in option for annotators to report any additional sounds beyond these categories. None of the write-in descriptions appeared with any regularity, suggesting that the eight categories are sufficient for describing the sounds present in these recordings.

These clips were split into train, validation, and test sets in proportions of 60%, 20%, and 20%. Classifier performance is evaluated using area under the receiver operating characteristic (AUC) within each class.

4.2. Exp 1: Manual Audio Set mapping

In order to produce predictions of our eight categories from the Audio Set model of [19], we first experimented with a manual mapping, combining multiple Audio Set labels into each event category. The following list shows mappings between the event categories and their corresponding Audio Set labels (separated by semicolons):

Tag	NPos	Bulbul	Manual	Audio Set	VGGish10	VGGish1
Wind	641	0.70	0.66	0.85 (gp)	0.90 (gp)	**0.91** (nn)
Cable noise	456	0.70	0.65	0.80 (rbf)	**0.87** (gp)	0.86 (gp)
Songbird	409	**0.86**	0.70	0.77 (gp)	0.83 (nn)	**0.86** (nn)
Running water	210	0.70	0.57	0.85 (gp)	**0.92** (nn)	0.89 (nn)
Water bird	196	0.65	0.59	0.74 (gp)	0.76 (nn)	**0.77** (rbf)
Insect	190	0.58	0.66	0.79 (nn)	**0.87** (lsvm)	0.82 (lsvm)
Rain	102	0.56	0.44	0.81 (rbf)	**0.85** (gp)	0.82 (gp)
Aircraft	28	0.66	0.52	0.78 (nn)	**0.86** (ab)	0.52 (gp)

Table 1. Labels applied to 1300 10-second clips, the total number positive examples annotated (NPos), and area under the ROC curve (AUC) results on our test set for bulbul, manual grouping of Audio Set tags (Manual), and classic machine learning models taking as input either Audio Set label predictions (Audio Set), averaged raw VGGish embeddings (VGGish10), or single VGGish embeddings (VGGish1). The best classifier was selected for each task on the validation set and is reported next to the result on the test set: lsvm: *Linear Support Vector Machine (SVM)*, rbf: *Radial Basis Function SVM*, nn: *Neural Network*, ab: *AdaBoost*, gp: *Gaussian process*

Songbird: Bird; Owl; Bird vocalization, bird call, bird song; Pigeon, dove; Coo; Chirp, tweet; Squawk; Bird flight, flapping wings; Gull, seagull; Chirp tone; Hoot

Water Bird: Duck; Goose; Quack; Frog; Croak; Caw

Insect: Fly, housefly; Insect; Bee, wasp, etc.; Buzz; Mosquito; Cricket; Rustle

Aircraft: Engine; Fixed-wing aircraft, airplane; Aircraft engine, Propeller, airscrew; Aircraft; Helicopter

Running Water: Waterfall; Waves, surf

Cable: Bang; Slap, smack; Whack, thwack; Smash, crash; Breaking; Knock; Tap; Thump, thud; Whip; Flap; Clip-clop

Wind: Wind; Howl

Rain: Rain; Raindrop; Rain on surface

Predictions for each event category were taken to be the maximum prediction confidence over all of the labels grouped into that category. Note that no label was included in more than one category.

Results for this approach are shown in the "Manual" column of Table 1. It was able to achieve AUC scores of 0.70 for Songbird, 0.66 for Insect and Wind, 0.65 for Cable noise, and 0.59 for Water bird.

4.3. Exp 2: Traditional classifiers on top of Audio Set labels

Results obtained using manual Audio Set mapping are lower than we expected, so we investigated the relationship between each Audio Set label and each event category individually in order to understand the model's behaviour. Figure 2 shows the AUC score on the test set for each category given the predictions of a single Audio Set label. For readability, only the 38 labels with the highest scores are shown. This heatmap includes many examples of labels successfully predicting categories that should not be related, e.g., the Animal label has a higher AUC in predicting the Rain category than it does for Songbird, Water bird, or Insect.

In order to address these issues of misalignment between descriptions and predictions, we trained several classic machine learning algorithms to predict our event categories from the predicted Audio Set labels. We choose following ML algorithms: nearest neighbors, linear support vector machine (SVM), radial basis function SVM, gaussian process, decision tree, random forest, multi-layer perceptron, and adaptive boosting. These classifiers used predictions over all Audio Set labels as input since we found that attempting to limit the labels hurt performance.

Results for these classifiers are shown in the "Audio Set" column of Table 1. For each event category, we select the classifier that performs best on the validation set and report its results on the test set in the table. Comparing the "Audio Set" column to the "Manual" column shows consistent improvements. For instance, the AUC score of Songbird increased by 7 percentage points (pp) from 0.70 to 0.77, Insect by 13 pp, and Water Bird by 15 pp.

4.4. Exp 3: Traditional classifiers on VGGish embeddings

Because of the general mismatch between the Audio Set labels and our event categories, we investigated whether these same classical machine learning models could predict our categories from the VGGish embeddings directly, without the intervening Audio Set labels. Thus, we trained classic machine learning algorithms on the raw and normalized VGGish embeddings. Since VGGish generates a 128-dimensional embedding vector for each second of audio, we experimented with two methods of combining ten of these vectors to make a prediction for a 10-second clip: averaging the embedding vectors and making predictions from the average, as well as concatenating the embedding vectors. We found that averaging worked better than concatenating and that the raw VGGish embeddings worked better than the scaled embeddings on the development set, so we use those.

Results from this approach are shown in the "VGGish10" column of Table 1 and show consistent improvement over the predictions based on the Audio Set labels. Comparing to the "Audio Set" column shows improvements of 8 pp for Insect and Aircraft, 7 pp for Cable noise and Running water, 6 pp for Songbirds, 5 pp for Wind, 4 pp for Rain, and 2 pp for Water birds.

We also experimented with weakly supervised classifiers of 1-second recordings. To do this, we divided each 10-second sample into 10 1-second samples, while assigning the labels of the original sample to all 10 samples. Then, we trained the classical ML models on those embeddings, and at inference time used the maximum over the 10 individual predictions for each category as the prediction for the 10-second clip. Results from this approach are shown in the "VGGish1" column of Table 1 and show that it performs similarly to "VGGish10". Notably, it achieves better results on the Songbird class using the Multi-layer Perceptron (MLP) classifier, bringing it to parity with the Bulbul model, which is specifically designed to recognize all birds. Table 1 shows that Bulbul does not identify Water birds well, achieving an AUC of 0.65 compared to our VGGish1 model's 0.77. Bulbul achieves a higher AUC on Wind, Cable Noise,

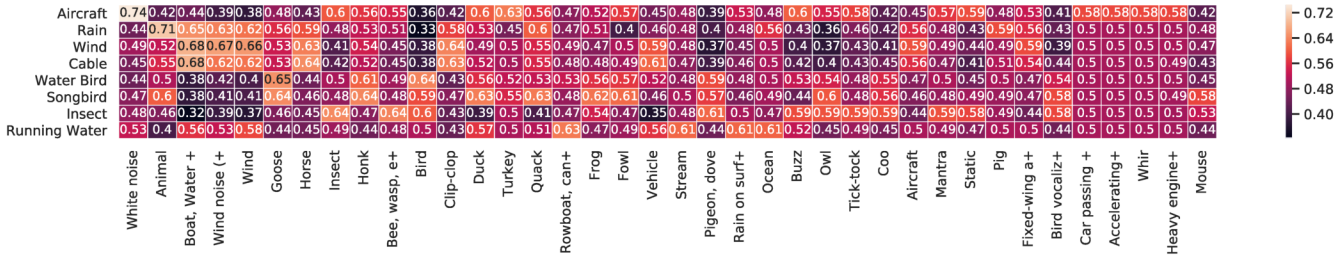

Fig. 2. Area under the ROC curve (AUC) score of each attention-based Audio Set classification label (columns) for predicting each of our event categories (rows). Long labels are truncated and are indicated with a "+" suffix. Rows and columns are sorted by their maximum value.

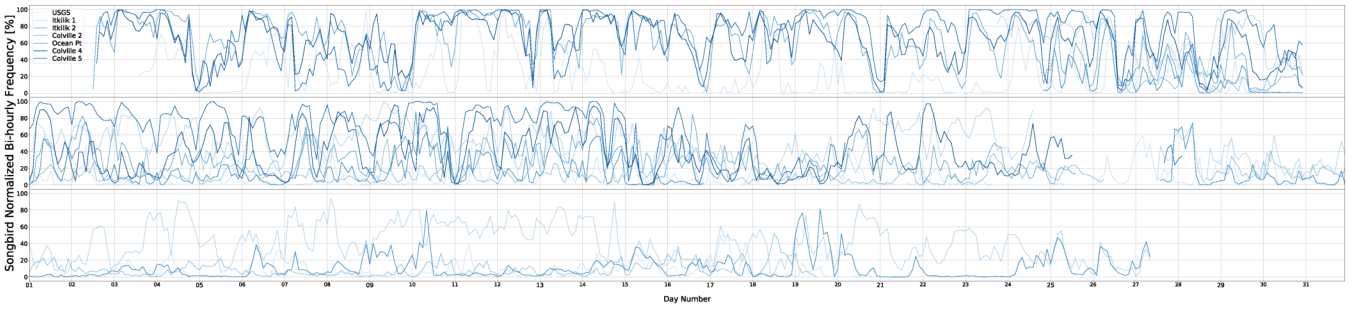

Fig. 3. Predictions of "Songbird" event category from Neural Network model trained on VGGish raw embeddings on all available data for 7 recording sites. Top: June, Middle: July, Bottom: August. Sites are listed in Legend and colored from most northerly ("USGS", Latitude 70.46°) to most southerly ("Colville 5", Latitude 69.89°), a distance of approximately 44 miles.

and Running water than Water birds. Note that Bulbul only makes a single prediction, and here we are evaluating this same prediction against each of our event categories.

This deficiency of Bulbul on Water bird prediction likely comes from the data that it was trained on, namely the datasets "freefield1010" (7,690 excerpts from field recordings around the world), "warblrb10k" (8,000 smartphone audio recordings from around the UK), and "BirdVox-DCASE-20k" (20,000 audio clips of flight calls collected from remote monitoring units near Ithaca, NY).

With this last experiment, we found a model that is performing well enough that we can use it with a certain confidence. We used it to predict Songbird events over the three months from the sites for which we have the longest duration of recordings. We visualized these predictions in Figure 3, which shows them at a 2-hour resolution.

Figure 3 shows a good amount of agreement between the predictions at different sites, suggests that the approach is relatively self-consistent. The one site that is less similar is "USGS", which is on the coast as opposed to the other sites, which are inland. Overall songbird detection drops substantially around July 16 at all of the inland sites, but continues through mid-August at the coastal site.

5. CONCLUSIONS AND FUTURE WORK

This paper examined the use of transfer learning to identify eight acoustic event categories of interest in recordings from an acoustic sensor network in Norther Alaska in the summer of 2016. We found that it was possible to identify all but one of the eight categories with AUC above 80% using classical machine learning models taking VGGish embeddings as input. This approach performed as well as Bulbul at detecting songbirds, which Bulbul was designed to detect, but was able to predict the other categories much more accurately. This approach also out-performed models based on manual or automatic grouping of Audio Set predictions from the attention-based model of Kong et al. [19]. The one category that was not recognized at above 80% AUC was Water birds. This might be because we grouped waterfowl together with shorebirds, making recognition more difficult for our systems.

These results show promise for analyzing soundscape recordings that we are currently collecting from a larger area of Northern Alaska at a larger number of sites. Future work will break these categories down into a finer granularity, potentially to the species level [31, 32, 33]. We plan to focus on utilizing these recordings and their analyses to identify important events in the phenology of the bird communities in these recording areas as well as measuring the characteristics of human-generated noise that might affect the phenology of caribou herds.

6. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation (NSF) grant OPP-1839185. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

7. REFERENCES

[1] J Richter-Menge, MO Jeffries, and E Osborne, "The arctic," in *Bulletin of the American Meteorological Society: State of the Climate in 2018*, Jessica Blunden and Derek S. Arndt, Eds., vol. 99, chapter 5, pp. S143–S168. 2018.

[2] Bryan C Pijanowski, Almo Farina, Stuart H Gage, Sarah L

Dumyahn, and Bernie L Krause, "What is soundscape ecology? an introduction and overview of an emerging new science," *Landscape ecology*, vol. 26, no. 9, pp. 1213–1232, 2011.

[3] Amandine Gasc, Jérôme Sueur, Sandrine Pavoine, Roseli Pellens, and Philippe Grandcolas, "Biodiversity sampling using a global acoustic approach: contrasting sites with microendemics in new caledonia," *PLoS One*, vol. 8, no. 5, pp. e65311, 2013.

[4] Wen Hu, Nirupama Bulusu, Chun Tung Chou, Sanjay Jha, Andrew Taylor, and Van Nghia Tran, "Design and evaluation of a hybrid sensor network for cane toad monitoring," *ACM TOSN*, vol. 5, no. 1, pp. 4, 2009.

[5] Almo Farina, *Soundscape ecology: principles, patterns, methods and applications*, Springer, 2013.

[6] Jason Wimmer, Michael Towsey, Paul Roe, and Ian Williamson, "Sampling environmental acoustic recordings to determine bird species richness," *Ecological Applications*, vol. 23, no. 6, pp. 1419–1428, 2013.

[7] Julia Shonfield and Erin Bayne, "Autonomous recording units in avian ecological research: current use and future applications," *Avian Conservation and Ecology*, vol. 12, no. 1, 2017.

[8] Dan Stowell and Mark D Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, pp. e488, 2014.

[9] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?," in *NIPS*, 2014, pp. 3320–3328.

[10] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "CNN architectures for large-scale audio classification," in *Proc. IEEE ICASSP*, 2017, pp. 131–135.

[11] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP*, 2017, pp. 776–780.

[12] Taylor R Stinchcomb, "Social-ecological soundscapes: examining aircraft-harvester-caribou conflict in arctic alaska," M.S. thesis, University of Alaska Fairbanks, 2017.

[13] Stephen R. Braund and Associates, "Nuiqsut caribou subsistence monitoring project: Results of year 6 hunter interviews and household harvest surveys," Tech. Rep., Prepared for ConocoPhillips Alaska, Inc., Anchorage, Alaska, 2015.

[14] Terry V. Callaghan, Margareta Johansson, Ross D. Brown, and Pavel Ya. Groisman et al, "The changing face of arctic snow cover: A synthesis of observed and projected changes," *AMBIO*, vol. 40, no. 1, pp. 17–31, Dec. 2011.

[15] Glen E. Liston and Christopher A. Hiemstra, "The Changing Cryosphere: Pan-Arctic Snow Trends (1979–2009)," *Journal of Climate*, vol. 24, no. 21, pp. 5691–5712, Jun. 2011.

[16] Ruth Y. Oliver, Daniel P. W. Ellis, Helen E. Chmura, Jesse S. Krause, Jonathan H. Pérez, Shannan K. Sweet, Laura Gough, John C. Wingfield, and Natalie T. Boelman, "Eavesdropping on the Arctic: Automated bioacoustics reveal dynamics in songbird breeding phenology," *Science Advances*, vol. 4, no. 6, Jun. 2018.

[17] Daniel P. W. Ellis, Xiaohong Zeng, and Josh H. McDermott, "Classifying soundtracks with audio texture features," in *Proc. IEEE ICASSP*, May 2011, pp. 5880–5883.

[18] Thomas Grill and Jan Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *Proc. EUSIPCO*, Aug. 2017, pp. 1764–1768.

[19] Qiuqiang Kong, Changsong Yu, Yong Xu, Turab Iqbal, Wenwu Wang, and Mark D. Plumbley, "Weakly labelled audioset tagging with attention neural networks," *IEEE Tr. Aud., Spch., & Lang. Proc.*, vol. 27, no. 11, pp. 1791–1802, Nov. 2019.

[20] Jérôme Sueur and Almo Farina, "Ecoacoustics: the ecological investigation and interpretation of environmental sound," *Biosemiotics*, vol. 8, no. 3, pp. 493–502, 2015.

[21] Jérôme Sueur, Almo Farina, Amandine Gasc, Nadia Pieretti, and Sandrine Pavoine, "Acoustic indices for biodiversity assessment and landscape investigation," *Acta Acustica united with Acustica*, vol. 100, no. 4, pp. 772–781, 2014.

[22] Zuzana Burivalova, Edward T Game, and Rhett A Butler, "The sound of a tropical forest," *Science*, vol. 363, no. 6422, pp. 28–29, 2019.

[23] Miguel A Acevedo, Carlos J Corrada-Bravo, Héctor Corrada-Bravo, Luis J Villanueva-Rivera, and T Mitchell Aide, "Automated classification of bird and amphibian calls using machine learning: A comparison of methods," *Ecological Informatics*, vol. 4, no. 4, pp. 206–214, 2009.

[24] Dan Stowell, Mike Wood, Yannis Stylianou, and Hervé Glotin, "Bird detection in audio: a survey and a challenge," in *Proc. MLSP*, 2016, pp. 1–6.

[25] Chenn-Jung Huang, Yi-Ju Yang, Dian-Xiu Yang, and You-Jia Chen, "Frog classification using machine learning techniques," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3737–3743, 2009.

[26] Ivan Himawan, Michael Towsey, Bradley Law, and Paul Roe, "Deep learning techniques for koala activity detection," *Proc. Interspeech*, pp. 2107–2111, 2018.

[27] Bernie Krause and Almo Farina, "Using ecoacoustic methods to survey the impacts of climate change on biodiversity," *Biological Conservation*, vol. 195, pp. 245–254, 2016.

[28] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[29] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[30] Dan Stowell, Michael D. Wood, Hanna Pamuła, Yannis Stylianou, and Hervé Glotin, "Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[31] Stefan Kahl, Fabian-Robert Stoter, Herve Goeau, Herve Glotin, Willem-Pier Vellinga, and Alexis Joly, "Overview of BirdCLEF 2019: Large-scale bird recognition in soundscapes," in *CLEF Working Notes*, 2019, p. 9.

[32] Stefan Kahl, Thomas Wilhelm-Stein, Holger Klinck, Danny Kowerko, and Maximilian Eibl, "A baseline for large-scale bird species identification in field recordings," in *CLEF Working Notes*, 2019, p. 4.

[33] Mario Lasseck, "Bird species identification in soundscapes," in *CLEF Working Notes*, 2019, p. 10.