# Feature Representations for Conservation Bioacoustics: Review and Discussion

**Irina Tolkova**
School of Engineering and Applied Sciences, Harvard University, USA
itolkova@g.harvard.edu

## Abstract

Acoustic analysis is becoming a key element of environmental monitoring for wildlife conservation. Passive acoustic recorders can document a variety of vocal animals over large areas and long time horizons, paving the path for machine learning algorithms to identify individual species, estimate abundance, and evaluate ecosystem health. However, such techniques rely on finding meaningful characterizations of calls and soundscapes, capable of capturing complex spatiotemporal, taxonomic, and behavioral structure. This article reviews existing methods for computing informative lower-dimensional features in the context of terrestrial passive acoustic monitoring, and discusses directions for further work.

## 1 Introduction

We are witnessing the sixth mass extinction, with staggering declines in species spanning the phylogenetic tree [Ceballos *et al.*, 2015; Pennisi, 2019]. These declines are understood to be driven by a collection of anthropogenic causes including climate change, deforestation, pollution, and poaching, compelling a crucial need for conservation efforts along with better understanding of the underlying ecological and biological systems [Grooten and Almond, 2018; Hanski, 2011]. To design and evaluate conservation programs, it is necessary to understand a species' population size, or at least estimate the change in biodiversity over time. Yet population estimation has traditionally required manual count surveys or catch-and-release operations by expert ecologists, assisted by statistical methods for data extrapolation [Taylor and Pollard, 2008; Sauer and Droege, 1990]. The high resource requirements, unreliability, and invasive nature of such surveys make it impractical to thoroughly monitor and understand the intertwined population dynamics of many species across an ecosystem, thereby limiting conservation efforts.

Fortunately, technological progress has driven the rise of several new methods in environmental monitoring. First, manual surveys are becoming replaced by networks of camera traps placed throughout an ecosystem [Burton *et al.*, 2015; Sollmann *et al.*, 2013]. Unlike a team of human observers, camera traps don't alarm wildlife and can actively monitor for a long time period. While this approach has proven effective for observing some animals – particularly large, elusive mammals – cameras can only capture a limited spatial area, and are not suitable for species which are small or consistently occluded [Trolliet *et al.*, 2014].

An increasingly popular alternative approach to study wild populations is through acoustics. Passive acoustic monitoring (PAM) can be used to identify individual species, study animal behavior, estimate population size, and evaluate overall biodiversity, and has been applied to birds, frogs, cetaceans, insects, elephants, bats, and others [Sugai *et al.*, 2019]. This article will review techniques for representing acoustic data within terrestrial soundscapes, with a focus on birdsong. I aim to give an algorithmic overview of feature extraction from the perspective of lower-dimensional "latent" spaces, discussing both traditional methods along with more recent optimization-based and deep-learning-based methods. Finally, the discussion section will touch on possible avenues for future work.

## 2 Overview

### 2.1 Tasks

In general, the primary aim of bio-acoustic studies is automatic detection of vocalizations within lengthy recordings, or species-level classification of calls or song. While interspecific differences in calls are challenging to characterize, it should be noted that even intraspecific classification is possible [Stowell *et al.*, 2019; Fox *et al.*, 2008]. Additionally, acoustic analysis has sparked interest as a method to monitor overall biodiversity. A variety of acoustic metrics have been proposed, such as spectral entropy, acoustic complexity, acoustic richness, normalized soundscape difference index, and others [Towsey *et al.*, 2014; Fuller *et al.*, 2015]. However, biodiversity indices are considered to have limited reliability due to high sensitivity to site- and survey- specific conditions [Gibb *et al.*, 2019].

### 2.2 Features and Latent Spaces

Generally, data analysis follows a pipeline with discrete steps: pre-processing (filtering, noise reduction, spectrogram calculation), segmentation (detection of calls within the recording), feature calculation, and training of a classifier over calls or spectrogram windows. But as the choice of features is often treated as a black box, the multitude of techniques for feature extraction, clustering, classification, and dimensionality reduction obscures the underlying objective of these methods – to find a numerical representation of the data which is able to reflect the classes or categories it is associated with. Any feature vector can then be considered a point in an implicit (*latent*) lower-dimensional space constructed by the algorithm. Viewing data analysis in this light allows for a more unified understanding of unsupervised and supervised learning – the former finds clusters in the latent space, the latter assigns labels to them. Furthermore, it allows to analyze complex semantic structure that can't be captured by a single label or classifier. For instance, calls may be characterized by type, species, sex, time of day, time of year, geography, or even individual; while a single classifier cannot simultaneously consider all of these factors, they can all be reflected within the structure of an appropriately-chosen latent space. This review will therefore interpret both feature extraction and dimensionality reduction as constructions of latent spaces, and review existing methods in four categories: traditional feature extraction, matrix factorization methods, graph-based mapping methods, and deep learning methods. In particular, discussion will focus on studies which demonstrate or analyze semantic structure.

### 2.3 Additional References

There have been many relevant surveys of bioacoustic literature. In particular, please see [Gibb *et al.*, 2019] for a general introduction and summary of passive acoustic monitoring within ecology; [Sugai *et al.*, 2019] for a large-scale analysis of prior work in terrestrial PAM; [Blumstein *et al.*, 2011] for a review of the deployment and analysis for terrestrial monitoring with microphone arrays; [Priyadarshani *et al.*, 2018] for a thorough survey of the birdsong classification pipeline; and [Stowell *et al.*, 2016] for an overview of bird detection in audio.

## 3 Algorithms for Feature Extraction

### 3.1 Traditional Methods

Traditional analysis of calls or song relies heavily on handcrafted features or spectral coefficients. In general, audio is analyzed by converting a time-domain signal to a spectrogram in the time-frequency domain through a discrete fourier transform (DFT). Since we do not perceive sound equally across all frequencies, but rather have logarithmic sensitivity, it is common for speech recognition algorithms to bin frequencies according to the mel-scale of human perception [Stevens *et al.*, 1937], a preprocessing step which is also commonly applied in the bioacoustic community. Common choices for hand-crafted features are peak frequency, highest and lowest frequency, call duration, number of harmonics, and energy; while these values are interpretable, they have limited representational ability [Priyadarshani *et al.*, 2018]. A ubiquitous alternative are mel-frequency cepstral coefficients (MFCCs) – an industry standard within speech processing and telecommunications [Muda *et al.*, 2010; Hasan *et al.*, 2004]. MFCCs are calculated by converting a signal to a mel-scale spectrogram, then calculating the coefficients which comprise the cepstrum (a variation on the Fourier spectrum) [Bogert, 1963]. Both of these approaches are commonly used and form baselines for more complex analysis [Somervuo *et al.*, 2006].

### 3.2 Matrix Factorization and Dictionary Learning

One branch of unsupervised feature learning is rooted in matrix decomposition methods. In particular, one of the oldest and most common methods for dimensionality reduction is principal component analysis (PCA), which finds an orthonormal basis for the directions of greatest variation within a dataset. Closely related methods include sparse PCA and robust PCA, which incorporate regularization terms [d'Aspremont *et al.*, 2005; Xu *et al.*, 2010]; independent component analysis (ICA), which instead optimizes for maximal independence of the basis vectors [Hyvärinen and Oja, 2000]; and non-negative matrix factorization (NMF), which constrains components to be non-negative [Lee and Seung, 2001]. Within acoustics, these methods have been employed for dictionary learning, such as to find the key elements (the "dictionary") that comprise urban sound [Bisot *et al.*, 2016], or the words and syllables that make up human speech [Tosic and Frossard, 2011; Bisot *et al.*, 2016; Jafari and Plumbley, 2011]. Likewise, studies in bioacoustics use dictionary learning for identifying core calls or song components [Ruiz-Muñoz *et al.*, 2018; Thakur *et al.*, 2018; Eldridge *et al.*, 2015; Seth *et al.*, 2018]. By decomposing a soundscape into a linear combination of components, the data can be represented in the low-dimensional space of coefficients.

### 3.3 Graph-Based Mapping Methods

Another branch of unsupervised learning constructs an optimal mapping of points to a low-dimensional space which faithfully represents their neighborhoods, configurations, or distances within the initial space. For example, Stochastic Neighbor Embedding (SNE) is a dimensionality reduction technique based on optimally preserving neighborhood probabilities of data points [Hinton and Roweis, 2003]. A slight variant on this approach, t-distributed Stochastic Neighbor Embedding (t-SNE) [Maaten and Hinton, 2008], improves the spread of points and simplifies the optimization procedure. Finally, Uniform Manifold Approximation and Projection (UMAP) is a similar algorithm grounded in manifold theory and topological data analysis [McInnes *et al.*, 2018]. UMAP accounts for preservation of global structure along with local structure, and is considered to be slightly advantageous to t-SNE. Both algorithms have been successfully applied to a number of large scientific datasets [Cao *et al.*, 2019; Becht *et al.*, 2019].

A number of bioacoustic studies have employed these methods for both feature extraction and visualization. For instance, [Parra-Hernández *et al.*, 2020] and [Valente *et al.*, 2019] showed that both methods clearly differentiated call type and geographic populations when applied to spectral/cepstral features of neotropical passerines and indri vocalizations. On a larger scale, [Sainburg *et al.*, 2019] analyzed the structure uncovered by UMAP for diverse vocalizations of 29 species, spanning songbirds, mice, primates, humans, and whales. In agreement with previous work, the study found that the latent space captured characteristics of individual identity, species identity, geographic variability, phonetic features, and acoustic categories, and allows for continuous analysis of calls and song composed of discrete elements. Moreover, this approach provides a visually informative view of differences in vocal structure across organisms.

### 3.4 Feature Extraction Through Neural Networks

In the last decade, convolutional neural networks (CNNs) have shown incredible success on image recognition tasks, and consequently have been adopted for bioacoustic classification. For instance, in 2015, none of the algorithms submitted to BirdCLEF – a workshop and big-data challenges for bird call classification – used neural networks, relying predominantly on MFCCs [Joly *et al.*, 2015]. In contrast, in 2019, all submissions featured CNN architectures, incorporating recent innovations such as attention, inception modules, and ensemble learning [Kahl *et al.*, 2019]. Furthermore, CNN-based software systems are being developed for large-scale use [Kahl, 2020].

While most deep learning systems are end-to-end – classification is learned directly over spectrogram inputs – there are a few techniques for understanding the underlying representation built by the network. First, neuron activations of a specific layer can themselves be considered features, albeit not necessarily lower-dimensional. [Sethi *et al.*, 2020] applied this approach to a network (VGGish) pre-trained on Google's AudioSet data, and utilized UMAP for further dimensionality reduction. The study found that these features were strongly structured by region, time of day, seasonality, and ecosystem type, showing that this latent space is able to capture meaningful patterns even without training on bioacoustic data. Knowledge of the soundscape embedding could be used to detect acoustic anomalies such as gunshots or chainsaw sounds, and extended to predict species occurrence, without relying on individualized classifiers.

Otherwise, dimensionality reduction through deep learning is closely associated with autoencoders (AEs): an architecture designed to "reconstruct" data by penalizing the input and output layers to be equal despite an informational bottleneck in the middle layers of the network, thereby optimizing for data compression. Many variants build on this idea, such as denoising AEs (which learn to reconstruct "clean" samples from "noisy" inputs) [Vincent *et al.*, 2008], sparse AEs (which are augmented with a regularization term) [Hosseini-Asl *et al.*, 2015], and variational AEs (VAEs, which define a generative model over the latent space) [Kingma and Welling, 2013]. A couple studies have utilized autoencoders to assist within a classification pipeline: [Narasimhan *et al.*, 2017] performed automatic segmentation and classification with an AE, while [Qiao *et al.*, 2020] learned lower-dimensional features prior to classification. Additionally, [Sainburg *et al.*, 2019] and [Goffinet *et al.*, 2019] analyzed latent spaces learned by VAEs for vocalizations of both wild species and laboratory animals. Similarly to Sainburg's findings with UMAP, both demonstrated strong structure across known data characteristics and accurate representation of vocal similarities and differences.

## 4 Discussion

Overall, many techniques for determining features have been developed to perform classification. By analyzing these features as points in a latent space, we can look for anticipated structure and evaluate the effectiveness of the feature extraction method, and even uncover new patterns in vocalization. While some studies have applied these methods in laboratory settings, there has been relatively little work in using latent spaces to understand calls and soundscapes in the wild (especially with autoencoder architectures). Moreover, most bioacoustic projects either characterize individual calls, or study soundscape-wide acoustic indices; learning lower-dimensional representations could yield an intermediate approach. For instance, by separating soundscapes into components which cannot be individually identified – such as crickets, frogs, flocked birdsong, and anthropogenic noise – we could gain valuable information about the overall intensity and change over time in these components, and design more robust and informative biodiversity metrics.

In comparison to camera traps, acoustic monitoring can be applied to a wider range of taxa, independent of body size or visual conditions – but can only provide information about vocal animals [Browning *et al.*, 2017]. In this degree, the two methods complement each other, providing an underutilized opportunity to develop more holistic ecosystem understanding through audiovisual features. Another source of information which can inform classification is metadata, which could allow to learn species distribution and movement patterns over large spatial and temporal scales. While contemporary classification methods do incorporate metadata, further work is needed to establish best practices for data fusion and analysis.

## 5 Conclusion

All in all, a diverse array of methods is used for characterizing sounds and soundscapes in conservation bioacoustics. There are also many directions for further work – particularly through an increased focus on latent spaces, rather than label-specific classification; in exploration of autoencoder architectures; in soundscape segmentation for informed biodiversity indices; and in audiovisual learning. Finally, while there have been great gains in analytical approaches for soundscape characterization, their use in practice has been limited: as of 2018, about 60% of studies in terrestrial bioacoustics relied on manual analysis [Sugai *et al.*, 2019]. This suggests that alongside improving classification accuracy, a greater emphasis should be placed on robustness, low sample complexity, usability, and communication of algorithms and software systems to potential users.

## Acknowledgments

This work was supported by the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (award number #1764269) and the Harvard Quantitative Biology Initiative. Thank you to Prof. Lakshminayaranan Mahadevan and Prof. Holger Klinck for encouraging my interest in biological acoustics.

## References


[Becht *et al.*, 2019] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.

[Bisot *et al.*, 2016] Victor Bisot, Romain Serizel, Slim Essid, and Gaël Richard. Acoustic scene classification with matrix factorization for unsupervised feature learning. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6445–6449. IEEE, 2016.

[Blumstein *et al.*, 2011] Daniel T Blumstein, Daniel J Mennill, Patrick Clemins, Lewis Girod, Kung Yao, Gail Patricelli, Jill L Deppe, Alan H Krakauer, Christopher Clark, Kathryn A Cortopassi, et al. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology*, 48(3):758–767, 2011.

[Bogert, 1963] Bruce P Bogert. The quefrency alanysis of time series for echoes; cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. *Time series analysis*, pages 209–243, 1963.

[Browning *et al.*, 2017] Ella Browning, Rory Gibb, Paul Glover-Kapfer, and Kate E Jones. Passive acoustic monitoring in ecology and conservation. 2017.

[Burton *et al.*, 2015] A Cole Burton, Eric Neilson, Dario Moreira, Andrew Ladle, Robin Steenweg, Jason T Fisher, Erin Bayne, and Stan Boutin. Wildlife camera trapping: a review and recommendations for linking surveys to ecological processes. *Journal of Applied Ecology*, 52(3):675–685, 2015.

[Cao *et al.*, 2019] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

[Ceballos *et al.*, 2015] Gerardo Ceballos, Paul R Ehrlich, Anthony D Barnosky, Andrés García, Robert M Pringle, and Todd M Palmer. Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science advances*, 1(5):e1400253, 2015.

[d'Aspremont *et al.*, 2005] Alexandre d'Aspremont, Laurent E Ghaoui, Michael I Jordan, and Gert R Lanckriet. A direct formulation for sparse pca using semidefinite programming. In *Advances in neural information processing systems*, pages 41–48, 2005.

[Eldridge *et al.*, 2015] Alice C Eldridge, Michael A Casey, Paola Moscoso, and Mika Peck. A new direction for soundscape ecology? toward the extraction and evaluation of ecologically-meaningful soundscape objects using sparse coding methods. *PeerJ PrePrints*, 3:e1407, 2015.

[Fox *et al.*, 2008] Elizabeth JS Fox, J Dale Roberts, and Mohammed Bennamoun. Call-independent individual identification in birds. *Bioacoustics*, 18(1):51–67, 2008.

[Fuller *et al.*, 2015] Susan Fuller, Anne C Axel, David Tucker, and Stuart H Gage. Connecting soundscape to landscape: Which acoustic index best describes landscape configuration? *Ecological Indicators*, 58:207–215, 2015.

[Gibb *et al.*, 2019] Rory Gibb, Ella Browning, Paul Glover-Kapfer, and Kate E Jones. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, 10(2):169–185, 2019.

[Goffinet *et al.*, 2019] Jack Goffinet, Richard Mooney, and John Pearson. Inferring low-dimensional latent descriptions of animal vocalizations. *bioRxiv*, page 811661, 2019.

[Grooten and Almond, 2018] M Grooten and RE Almond. A (eds.)(2018) living planet report—2018: Aiming higher. *Gland, Switzerland: WWF*, 2018.

[Hanski, 2011] Ilkka Hanski. Habitat loss, the dynamics of biodiversity, and a perspective on conservation. *Ambio*, 40(3):248–255, 2011.

[Hasan *et al.*, 2004] Md Rashidul Hasan, Mustafa Jamil, MGRMS Rahman, et al. Speaker identification using mel frequency cepstral coefficients. *variations*, 1(4), 2004.

[Hinton and Roweis, 2003] Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864, 2003.

[Hosseini-Asl *et al.*, 2015] Ehsan Hosseini-Asl, Jacek M Zurada, and Olfa Nasraoui. Deep learning of part-based representation of data using sparse autoencoders with nonnegativity constraints. *IEEE transactions on neural networks and learning systems*, 27(12):2486–2498, 2015.

[Hyvärinen and Oja, 2000] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[Jafari and Plumbley, 2011] Maria G Jafari and Mark D Plumbley. Fast dictionary learning for sparse representations of speech signals. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):1025–1031, 2011.

[Joly *et al.*, 2015] Alexis Joly, Hervé Goëau, Hervé Glotin, Concetto Spampinato, Pierre Bonnet, Willem-Pier Vellinga, Robert Planqué, Andreas Rauber, Simone Palazzo, Bob Fisher, et al. Lifeclef 2015: multimedia life species identification challenges. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 462–483. Springer, 2015.

[Kahl *et al.*, 2019] Stefan Kahl, Fabian-Robert Stöter, Hervé Goëau, Hervé Glotin, Robert Planque, Willem-Pier Vellinga, and Alexis Joly. Overview of birdclef 2019: large-scale bird recognition in soundscapes. 2019.

[Kahl, 2020] M Sc Stefan Kahl. Identifying birds by sound: Large-scale acoustic event recognition for avian activity monitoring. 2020.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Lee and Seung, 2001] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[McInnes *et al.*, 2018] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[Muda *et al.*, 2010] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010.

[Narasimhan *et al.*, 2017] Revathy Narasimhan, Xiaoli Z Fern, and Raviv Raich. Simultaneous segmentation and classification of bird song using cnn. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 146–150. IEEE, 2017.

[Parra-Hernández *et al.*, 2020] Ronald M Parra-Hernández, Jorge I Posada-Quintero, Orlando Acevedo-Charry, and Hugo F Posada-Quintero. Uniform manifold approximation and projection for clustering taxa through vocalizations in a neotropical passerine (rough-legged tyrannulet, phyllomyias burmeisteri). *Animals*, 10(8):1406, 2020.

[Pennisi, 2019] Elizabeth Pennisi. Billions of north american birds have vanished. *Science*, 365(6459):1228–1229, 2019.

[Priyadarshani *et al.*, 2018] Nirosha Priyadarshani, Stephen Marsland, and Isabel Castro. Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology*, 49(5):jav–01447, 2018.

[Qiao *et al.*, 2020] Yu Qiao, Kun Qian, and Ziping Zhao. Learning higher representations from bioacoustics: A sequence-to-sequence deep learning approach for bird sound classification. 2020.

[Ruiz-Muñoz *et al.*, 2018] José Francisco Ruiz-Muñoz, Zeyu You, Raviv Raich, and Xiaoli Z Fern. Dictionary learning for bioacoustics monitoring with applications to species classification. *Journal of Signal Processing Systems*, 90(2):233–247, 2018.

[Sainburg *et al.*, 2019] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Latent space visualization, characterization, and generation of diverse vocal communication signals. *bioRxiv*, page 870311, 2019.

[Sauer and Droege, 1990] John R Sauer and Sam Droege. *Survey designs and statistical methods for the estimation of avian population trends*. US Department of the Interior, Fish and Wildlife Service, 1990.

[Seth *et al.*, 2018] Harshita Seth, Rhythm Bhatia, and Padmanabhan Rajan. Feature learning for bird call clustering. In *2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS)*, pages 72–76. IEEE, 2018.

[Sethi *et al.*, 2020] Sarab S Sethi, Nick S Jones, Ben D Fulcher, Lorenzo Picinali, Dena Jane Clink, Holger Klinck, C David L Orme, Peter H Wrege, and Robert M Ewers. Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. *Proceedings of the National Academy of Sciences*, 117(29):17049–17055, 2020.

[Sollmann *et al.*, 2013] Rahel Sollmann, Azlan Mohamed, Marcella J Kelly, et al. Camera trapping for the study and conservation of tropical carnivores. *Raffles Bulletin of Zoology*, 28:21–42, 2013.

[Somervuo *et al.*, 2006] Panu Somervuo, Aki Harma, and Seppo Fagerlund. Parametric representations of bird sounds for automatic species recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):2252–2263, 2006.

[Stevens *et al.*, 1937] Stanley Smith Stevens, John Volkmann, and Edwin B Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[Stowell *et al.*, 2016] Dan Stowell, Mike Wood, Yannis Stylianou, and Hervé Glotin. Bird detection in audio: a survey and a challenge. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2016.

[Stowell *et al.*, 2019] Dan Stowell, Tereza Petrusková, Martin Šálek, and Pavel Linhart. Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions. *Journal of the Royal Society Interface*, 16(153):20180940, 2019.

[Sugai *et al.*, 2019] Larissa Sayuri Moreira Sugai, Thiago Sanna Freire Silva, José Wagner Ribeiro Jr, and Diego Llusia. Terrestrial passive acoustic monitoring: review and perspectives. *BioScience*, 69(1):15–25, 2019.

[Taylor and Pollard, 2008] Sandra L Taylor and Katherine S Pollard. Evaluation of two methods to estimate and monitor bird populations. *PLoS One*, 3(8):e3047, 2008.

[Thakur *et al.*, 2018] Anshul Thakur, Vinayak Abrol, Pulkit Sharma, and Padmanabhan Rajan. Deep convex representations: Feature representations for bioacoustics classification. In *INTERSPEECH*, pages 2127–2131, 2018.

[Tosic and Frossard, 2011] Ivana Tosic and Pascal Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2):27–38, 2011.

[Towsey *et al.*, 2014] Michael Towsey, Jason Wimmer, Ian Williamson, and Paul Roe. The use of acoustic indices to determine avian species richness in audio-recordings of the environment. *Ecological Informatics*, 21:110–119, 2014.

[Trolliet *et al.*, 2014] Franck Trolliet, Cédric Vermeulen, Marie-Claude Huynen, and Alain Hambuckers. Use of camera traps for wildlife studies: a review. *Biotechnologie, Agronomie, Société et Environnement*, 18(3):446–454, 2014.

[Valente *et al.*, 2019] Daria Valente, Chiara De Gregorio, Valeria Torti, Longondraza Miaretsoa, Olivier Friard, Rose Marie Randrianarison, Cristina Giacoma, and Marco Gamba. Finding meanings in low dimensional structures: Stochastic neighbor embedding applied to the analysis of indri indri vocal repertoire. *Animals*, 9(5):243, 2019.

[Vincent *et al.*, 2008] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[Xu *et al.*, 2010] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. In *Advances in neural information processing systems*, pages 2496–2504, 2010.