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 Droge, 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 hand-crafted 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-Asli 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. Lakshminarayanan Mahadevan and Prof. Holger Klinck for encouraging my interest in biological acoustics.
References


[Trolliet et al., 2014] Franck Trolliet, Cédric Vermeulen, Marie-Claude Huynen, and Alain Hambuckers. Use of

