# Detection and Classification Methods for Animal Sounds

8

Julie N. Oswald, Christine Erbe, William L. Gannon, Shyam Madhusudhana, and Jeanette A. Thomas

## 8.1 Introduction

Researchers have a natural tendency to classify biological systems into categories. For example, organisms can be classified based on biome, ecosystem, taxon, phylogeny, niche, demographic class, behavior type, etc., and this allows complex systems to be organized. Categorization also can make recognition of patterns easier and assist in understanding the ways in which biological systems work. Classification provides a convenient method for comparing features, making systematic measurements, testing hypotheses, and performing statistical analyses.

Bioacousticians have categorized sounds produced by animals for decades, and new methods for classification continue to be developed (Horn and Falls 1996; Beeman 1998). Animals produce many different types of sounds that span orders of magnitude along the dimensions of time, frequency, and amplitude. For example, the repertoire of marine mammal acoustic signals includes broadband echolocation clicks as short as 10 μs in duration and with energy up to 200 kHz, as well as narrowband tonal sounds as low as 10–20 Hz, lasting more than10 s. Song birds and some species of baleen whales arrange individual sounds into patterns called song and repeat these patterns for hours or days. Some mammal species produce distinctive, stereotyped sounds (e.g., chipmunks, dogs, and blue whales), while others produce signals with high variability (e.g., mimicking birds, primates, and dolphins).

Because animals produce so many different types of sounds, developing algorithms to detect, recognize, and classify a wide range of acoustic signals can be challenging. In the past, detection and classification tasks were performed by an experienced bioacoustician who listened to the sounds and visually reviewed spectrographic displays (e.g., for birds by Baptista and Gaunt 1997; chipmunks by Gannon and Lawlor 1989; baleen whales by Stafford et al. 1999; and delphinids by Oswald et al. 2003). Before the advent of digital signal-analysis, data were

Jeanette A. Thomas (deceased) contributed to this chapter while at the Department of Biological Sciences, Western Illinois University-Quad Cities, Moline, IL, USA

J. N. Oswald (✉)
Scottish Oceans Institute, University of St Andrews, St Andrews, Fife, UK
e-mail: jno@st-andrews.ac.uk

C. Erbe
Centre for Marine Science & Technology, Curtin University, Perth, WA, Australia
e-mail: c.erbe@curtin.edu.au

W. L. Gannon
Department of Biology and Graduate Studies, Museum of Southwestern Biology, University of New Mexico, Albuquerque, NM, USA
e-mail: wgannon@unm.edu

S. Madhusudhana
K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA
e-mail: shyamm@cornell.edu


C. Erbe, J. A. Thomas (eds.), *Exploring Animal Behavior Through Sound: Volume 1*,
https://doi.org/10.1007/978-3-030-97540-1_8

analyzed while enduring the acrid smell of etched Kay Sona-Graph paper and piles of 8-s printouts removed from a spinning recording drum littering laboratory tables and floors. Output from a long-duration sound had to be spliced together (see Chap. 1). Many bioacoustic studies generated an enormous amount of data, which made this manual review process at best inefficient, and at worst impossible to accomplish.

For decades, scientists have worked to automate the process of detecting and classifying sounds into categories or types. Automated classification involves three main steps: (1) detection of potential sounds of interest, (2) extraction of relevant acoustic characteristics (or, features) from these sounds, and (3) classification of these sounds as produced by a particular species, sex, age, or individual. Methods for the automated detection of sounds have progressed quickly with technological advances in digital recording (see Chap. 2). Likewise, the extraction of sound variables useful in analysis has expanded with an increasing amount of information provided by new technology. For instance, where features such as maximum frequency or time between sounds originally were measured manually off sonagraph paper, devices today allow for measuring these, and many more variables, automatically or semi-automatically using computer software. Now, derived variables, such as time difference between individual signal elements, frequency modulation, running averages of sound frequency, and harmonic structure can be easily obtained for classifying the sounds in a repertoire.

Some of the earliest methods used for automated detection and classification included energy threshold detectors (e.g., Clark 1980) and matched filters (e.g., Freitag and Tyack 1993; Stafford et al. 1998; Dang et al. 2008; Mankin et al. 2008). These methods were used to detect and classify simple, stereotypical sounds produced by species such as the Asian longhorn beetle (*Anoplophora glabripennis*), cane toads (*Rhinella marina*), blue whales (*Balaenoptera* spp.), and fin whales (*Balaenoptera physalus*). Once sounds are detected, they can be organized into groups, or classified, based on selected acoustic characteristics. For example, development of methods for detection and automated signal processing of bat sounds led to a variety of automated, off-the-shelf, ready-to-deploy bat detectors that detect and classify sounds by species (Fenton and Jacobson 1973; Gannon et al. 2004). These detectors can be very useful in addressing biological or management issues in ecology, evolution, and impact mitigation. While the accuracy and robustness of automated approaches are always a matter of concern (Herr et al. 1997; Parsons et al. 2000), modern techniques promise much improved recognition performances that could rival manual analyses (e.g., Brown and Smaragdis 2009).

Multivariate statistical methods can be powerful for classification of sounds produced by species with variable vocal repertoires because they can identify complex relationships among many acoustic features (see Chap. 9). With the advent of powerful personal computers in the 1980s and 1990s, the use of multivariate techniques became popular for classifying bird sounds (e.g., Sparling and Williams 1978; Martindale 1980a, b). Since then, enormous effort has been expended to develop these and other automatic methods for the detection of sounds produced by many taxa and their classification into discrete categories, such as species, population, sex, or individual.

These days, there are applications (apps) for smartphones that use advanced algorithms to automatically detect and recognize sounds. For example, the *BirdNET* app detects and classifies bird song—similar to the *Shazam* app for music—and provides a listing of the top-ranked matching species. It includes almost 1000 of the most common species of North America and Europe. A similar app, *Song Sleuth*, recognizes songs of nearly 200 bird species likely to be heard in North America and also provides references for species identification, such as the David Sibley Bird Reference (Sibley 2000), allowing the user to "dig into" the bird's biology and conservation needs.

In this chapter, we present an overview of methods for detection and classification of sounds along with examples from different taxa. No single method is appropriate for every research

project and so the strengths and weaknesses of each method are summarized to help guide decisions on which methods are better suited for particular research scenarios. Because algorithms for statistical analyses, automated detection, and computer classification of animal sounds are advancing rapidly, this is not a comprehensive overview of methods, but rather a starting point to stimulate further investigations.

## 8.2 Qualitative Naming and Classification of Animal Sounds

Prior to computer-assisted detection and classification of animal sounds, bioacousticians used various qualitative methods to categorize sounds.

### 8.2.1 Onomatopoeic Names

Frequently, researchers describe and name animal sounds based on their perception of the sound and thus based on their own language. This approach has been common in the study of terrestrial animals (in particular, birds) and marine mammals (in particular, pinnipeds and mysticetes). Researchers also have given onomatopoeic names to sounds. These are names that phonetically resemble the sound they describe. For example, the sounds of squirrels and chipmunks have been described as barks, chatters, chirps, and growls. The primate literature is also rich in these sorts of sound descriptions (e.g., the hack sequences and boom-hack sequences described for Campbell's monkeys, *Cercopithecus campbelli*; Ouattara et al. 2009). Bioacousticians studying humpback whales (*Megaptera novaeangliae*) have described a repertoire of sounds including barks, bellows, chirps, cries, croaks, groans, growls, grumbles, horns, moans, purrs, screams, shrieks, sighs, sirens, snorts, squeaks, thwops, trumpets, violins, wops, and yaps (Dunlop et al. 2007, 2013). While it is potentially convenient for researchers within a group to discuss sounds this way, it is more difficult for others, and perhaps impossible for foreign-language speakers to recognize the sound type. An example of this difficulty in describing a sound is the ubiquitous rooster crow, which can be described by a US citizen as "cock-a-doodle-doo" and by a German citizen as "kikeriki". Roosters make the same sound, no matter in which country they live, yet their single sound has been named so differently, as has the bark of dogs (Fig. 8.1). Of course, onomatopoeic naming of sounds also fails when the sounds are outside of the human hearing range.

If the above was not confusing enough, bird calls have been described using onomatopoeic phrases. For example, the song of a white-throated sparrow (*Zonotrichia albicollis*) has been described in Canada as sounding like "O sweet Canada Canada Canada" and in New England, USA, as "Old Sam Peabody Peabody Peabody." Another example is the barred owl (*Strix varia*), which hoots "Who cooks for you? Who cooks for you all?".

### 8.2.2 Naming Sounds Based on Animal Behavior

Researchers sometimes name sounds based on observed and interpreted animal behavior. For example, the various echolocation signals described for insectivorous bats have been named "search clicks" (i.e., slow and regular clicks) while pursuing insect prey and "terminal feeding buzz" (i.e., accelerated click trains) during prey capture (Griffin et al. 1960). The bird and mammal literature is replete with sounds named for a behavior, such as the begging call of nestling chicks (Briskie et al. 1999; Leonard and Horn 2001), the contact call for isolated young (Kondo and Watanabe 2009), and the alarm call warning of a nearby predator (Zuberbuhler et al. 1999; Gill and Bierema 2013). In some cases, the function of sounds has been studied in detail, which justifies using their function in the name. Examples are feeding buzzes in echolocation or alarm calls in primates. However, naming sounds according to behavior can be misleading because a sound can be associated with several contexts. Names based on the associated behavior should really only be used after detailed studies of context-specificity of the calls in question.

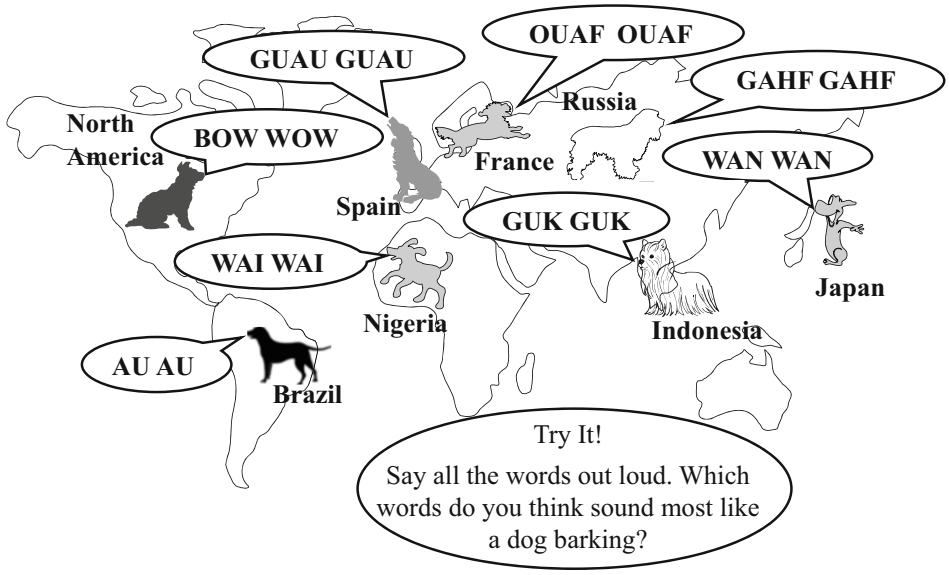


**Fig. 8.1** Dogs speak out. Labels used for dog barks in different countries

### 8.2.3 Naming Sounds Based on Mechanism of Sound Production

Some bioacousticians identify and classify sounds based on the mechanism of sound production. For example, one syllable in insect song corresponds to a single to- and fro-movement of a stridulatory anatomy or one cycle of a forewing opening and closing in the field cricket (*Gryllus* spp.). McLister et al. (1995) defined a note in chorusing frogs as the sound unit produced during a single expiration. Classifying sound types by their mode of production perhaps is less ambiguous and unequivocal, but there are limited data on the mechanisms of sound production in many animals.

### 8.2.4 Naming Sounds Based on Spectro-Temporal Features

An alternative, but not necessarily better, way of naming sounds is based on their spectro-temporal features. For instance, in distinguishing two morphologically similar species of bats, *Myotis californicus* is referred to as a "50-kHz bat" and *M. ciliolabrum* as a "40-kHz bat," which describes the terminal frequency of the downsweep of their ultrasonic echolocation signals (Gannon et al. 2001). Under water, the most common sound recorded from southern right whales (*Eubalaena australis*) is a 1–2 s frequency-modulated (FM) upsweep from about 50–200 Hz, commonly recorded with overtones, and referred to in the literature as the upcall (Fig. 8.2; Clark 1982). Antarctic blue whales (*Balaenoptera musculus intermedia*) produce a Z-call, which consists of a 10-s constant frequency (also called constant-wave, CW) sound at 28 Hz, followed by a rapid FM downsweep to 18 Hz, where the sound continues for another 15-s CW component (Rankin et al. 2005).

While the measurement of features from spectrograms and waveforms can be expected to be more objective than onomatopoeic or functional naming, the appearance of a spectrogram, and thus the measurements made, depend on characteristics of the recording system, the time and frequency settings of the analysis algorithm, and analysis algorithm used. This can make sounds look rather different at various scales and therefore lead to inconsistent classification.

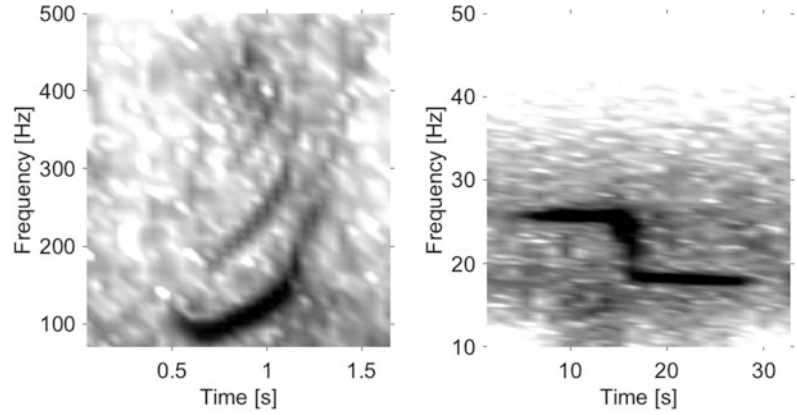


**Fig. 8.2** Spectrograms of southern right whale "upcall" (left; sampling frequency $f_s = 12$ kHz, Fourier window length $NFFT = 1200$, 50% overlap, Hann window) and Antarctic blue whale "Z-call" (right; $f_s = 6$ kHz, $NFFT = 16384$, 50% overlap, Hann window) recorded off southern Australia (Erbe et al. 2017)

An example of the confusion that can arise from different representations of sound is the boing sound made by minke whales (*Balaenoptera acutorostrata*), which was given an onomatopoeic name. In spectrograms, the boing might look like an FM sound (Fig. 8.3a), however, it is actually a series of rapid pulses (Rankin and Barlow 2005), similar to burst-pulse sounds produced by odontocetes (e.g., Wellard et al. 2015). As another example, the bioduck sound made by Antarctic minke whales (*Balaenoptera bonaerensis*) got its name because it resembles a duck's quack to human listeners (Risch et al. 2014). A spectrogram of the bioduck sound appears as a series of pulses; however, each pulse actually is a 0.3-s FM downswept tone from 300 to 100 Hz (Fig. 8.3b). As if this was not enough in terms of interesting sounds and odd names, dwarf minke whales produce the so-called star-wars sound, which is composed of a series of pulses with varying pulse rates (Gedamke et al. 2001). The different pulse rates make this sound appear as a mixture of broadband pulses and FM sounds in spectrograms, depending on the spectrogram settings (Fig. 8.3c). The sound name presumes the reader is familiar with the soundtrack of an American movie from the 1970s.

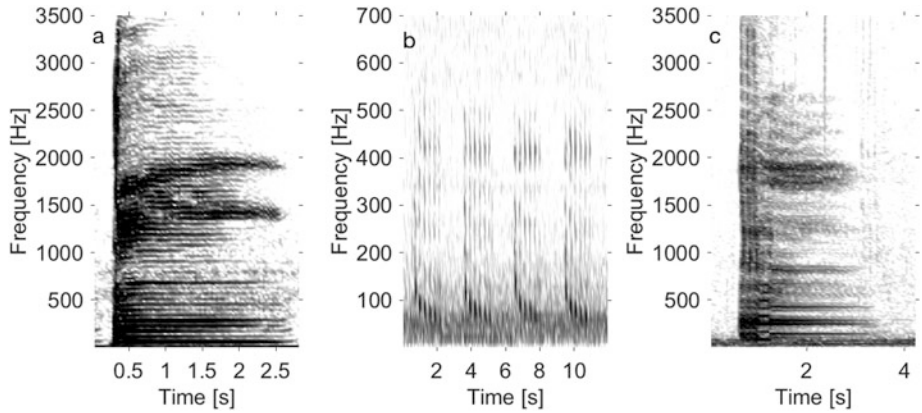


**Fig. 8.3** Spectrograms of the dwarf minke whale boing (**a** $f_s = 16$ kHz, $NFFT = 1024$, 50% overlap, Hann window), the Antarctic minke whale bioduck sound (**b** $f_s = 96$ kHz, $NFFT = 8192$, 50% overlap, Hann window), and the dwarf minke whale star-wars sound (**c** $f_s = 44$ kHz, $NFFT = 4096$, 50% overlap, Hann window). Recordings **a** and **b** from Erbe et al. (2017), **c** from Gedamke et al. (2001)

### 8.2.5 Naming Sounds Based on Human Communication Patterns

The term "song" is perhaps the best-known example of using human communication labels in the description of animal sounds. The word "song" may be used to simply indicate long-duration displays of a specific structure. Songs of insects and frogs are relatively simple sequences, consisting of the same sound repeated over long periods of time. The New River tree frog (*Trachycephalus hadroceps*), for example, produces nearly 38,000 calls in a single night (Starnberger et al. 2014). Many frogs use trilling notes in mate attraction, which has been described as song, but switch to a different vocal pattern in aggressive territorial displays (Wells 2007). In some frog songs, different notes serve different purposes, with one type of note warding off competing males, and another attracting females. In birds and mammals, songs are often more complex, consisting of several successive sounds in a recognizable pattern. They appear to be used primarily for territorial defense or mate attraction (Bradbury and Vehrencamp 2011). Our statements in this chapter show one way to describe calls and songs in animals; however, it is important to note that borrowing terminology from human communication when studying animals can lead to confusion. The terms we discuss here are not well defined and are used differently by different authors. Make sure to pay close attention to these definitions when reading literature about animal communication.

Some ornithologists have used human-language properties further to describe the structure of bird song. Song may be broken down into phrases (also called motifs). Each phrase is composed of syllables, which consist of notes (or elements, the smallest building blocks; Catchpole and Slater 2008). Notes, syllables, and phrases are identified and defined based on their repeated occurrence. An entire taxon of birds (songbirds, Order Passeriformes) has been designated by ornithologists because of their use of these elaborate sounds for territorial defense and/or mate attraction. Birds of this taxon usually use sets of sounds that are repeated in an organized structure. In many species, males produce such songs continuously for several hours each day, producing thousands of songs in each performance. In the bird song literature, songs are distinguished from calls by their more complex and sustained nature, species-typical patterns, or syntax that governs their combination of syllables and notes into a song. Songs are under the influence of reproductive hormones and associated with courtship (Bradbury and Vehrencamp 2011). Bird song can vary geographically and over time (e.g., Fig. 8.4; Camacho-Alpizar et al. 2018). In contrast, calls are typically acoustically simple and serve non-reproductive, maintenance functions, such as coordination of parental duties, foraging, responding to threats of predation, or keeping members of a group in contact (Marler 2004).

Several terrestrial mammals have been reported to sing. For instance, adult male rock hyraxes (*Procavia capensis*) engage throughout most of the year in rich and complex vocalization behavior that is termed singing (Koren et al. 2008). These songs are complex signals and are composed of multiple elements (chucks, snorts, squeaks, tweets, and wails) that encode the identity, age, body mass, size, social rank, and hormonal status of the singer (Koren and Geffen 2009, 2011). Holy and Guo (2005) described ultrasonic sounds from male laboratory mice (*Mus musculus*) as song. Von Muggenthaler et al. (2003) reported that Sumatran rhinoceros (*Dicerorhinus sumatrensis*) produce a song composed of three sound types: eeps (simple short signals, 70 Hz–4 kHz), humpback whale like sounds (100 Hz–3.2 kHz, varying in length, only produced by females), and whistle blows (loud, 17 Hz–8 kHz vocalizations followed by a burst of air with strong infrasonic content). Clarke et al. (2006) described the syntax and meaning of wild white-handed gibbon (*Hylobates lar*) songs.

Among marine mammals, blue, bowhead (*Balaena mysticetus*), fin, humpback, minke, and right whales, Weddell seals (*Leptonychotes weddellii*), harbor seals (*Phoca vitulina*), and

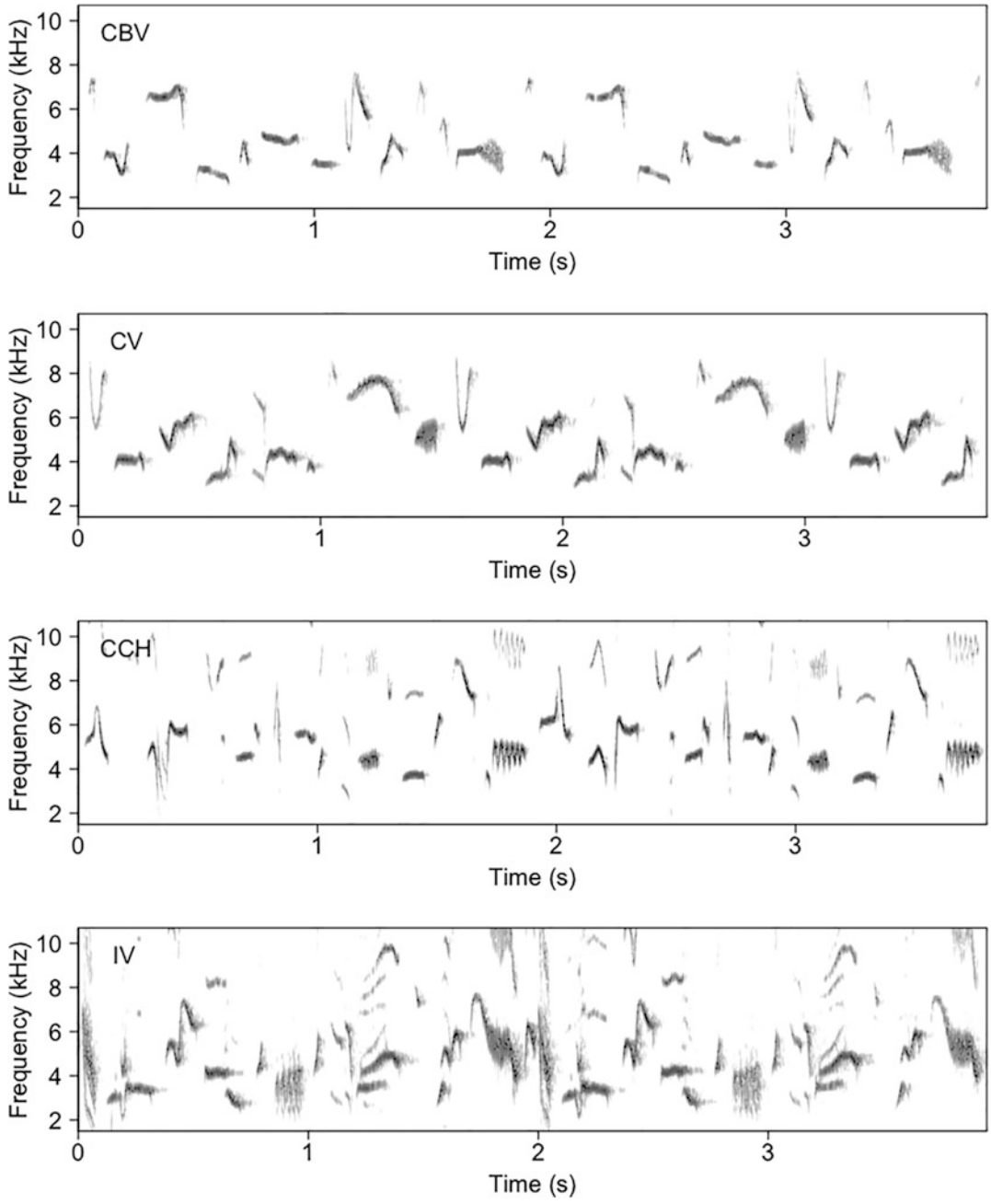


**Fig. 8.4** Geographic variation in birdsong. These spectrograms show a portion of song from Timberline wrens (*Thryorchilus browni*) recorded at four locations in Costa Rica (CBV = Cerro Buena Vista, CV = Cerro Vueltas, CCH = Cerro Chirripó, IV = Irazú Volcano) (Camacho-Alpizar et al. 2018). © Camacho-Alpizar et al.; https://doi.org/10.1371/journal.pone.0209508. Licensed under CC BY 4.0; https://creativecommons.org/licenses/by/4.0/

walrus (*Odobenus rosmarus*) have all been reported to sing (Payne and Payne 1985; Sjare et al. 2003; McDonald et al. 2006; Stafford et al. 2008; Oleson et al. 2014; Crance et al. 2019). The songs of blue, bowhead, fin, minke, and right whales are simple compared to those of the humpback whale and little is known about the behavioral context of song in any marine mammal species besides the humpback whale. Humpback whales are well-known for their long, elaborate songs. These songs are composed of themes consisting of repetitions of phrases made up of patterns of units similar to syllables in bird song (Fig. 8.5; Payne and Payne 1985; Helweg et al. 1998). Winn and Winn (1978) suggested that only male baleen whales sing, as a means of reproductive display. Sjare et al. (2003) reported that Atlantic walrus produce two main songs: the coda song and the diving vocalization song that differ by their pattern of knocks, taps, and bell sounds.

Song production does not exclude the emission of non-song sounds and most singing species likely emit both. The non-song sounds of humpback and pygmy blue whales (*Balaenoptera musculus brevicauda*), for example, have been cataloged (e.g., Recalde-Salas et al. 2014, 2020). Some song units may resemble non-song sounds.

Whether sounds are part of song or not, their detection and classification can be challenging when repertoires are large and possibly variable across time and space. Humpback whale songs, for example, vary by region and year (Cerchio et al. 2001; Payne and Payne 1985). Characterizing and describing the structure of song can be a difficult task even for the experienced bioacoustician. With the assistance of computer analysis tools, sound detection and classification may be more efficient.

## 8.3 Detection of Animal Sounds

The problem to be solved may seem simple. For example, a bioacoustician deployed an autonomous recorder in the field for a month, and after recovery of the gear, downloaded all data in the laboratory and now wants to pick all frog calls recorded in order to study the mating behavior of this species. Listening to the first few minutes of recording, the bioacoustician can easily hear the target species, but there are calls every few seconds—too many to pick by hand. So, the scientist looks for software tools to help detect all frog signals, and potentially sort them based on their acoustic features. The first step, signal detection, is discussed in Sect. 8.3; the second step, signal classification, is discussed in Sect. 8.4.

Automated signal detectors work by common principles. The raw input data are the ideally calibrated time series of pressure recorded with a microphone in air or hydrophone in water. There might be one or more pre-processing steps to filter or Fourier transform the data in successive time windows (see Chap. 4). The pre-processed time series is then fed into the detector, which computes a specific quantity from the acoustic data. This may be instantaneous energy, energy within a specified time window, entropy, or a correlation coefficient, as a few examples. Then, a detection threshold is applied. If the quantity exceeds the threshold, the signal is deemed present, otherwise not.

The threshold is commonly computed the following way:

$$E_{\text{th}} = \bar{E} + \gamma \sigma_E$$

where $E$ symbolizes the chosen quantity (e.g., energy), $\bar{E}$ is its mean value computed over a long time window (e.g., an entire file), $\sigma_E$ is the standard deviation, and $\gamma$ is a multiplier (integer or real). Setting a high threshold will result in only the strongest signals being detected and weaker ones being missed. Setting a low threshold will result in many false alarms, which are not signals. By varying $\gamma$, the ideal threshold may be found and the performance of the detector may be assessed (see Sect. 8.3.6).

### 8.3.1 Energy Threshold Detector

One of the most common methods for detecting animal sounds from recordings is to measure the

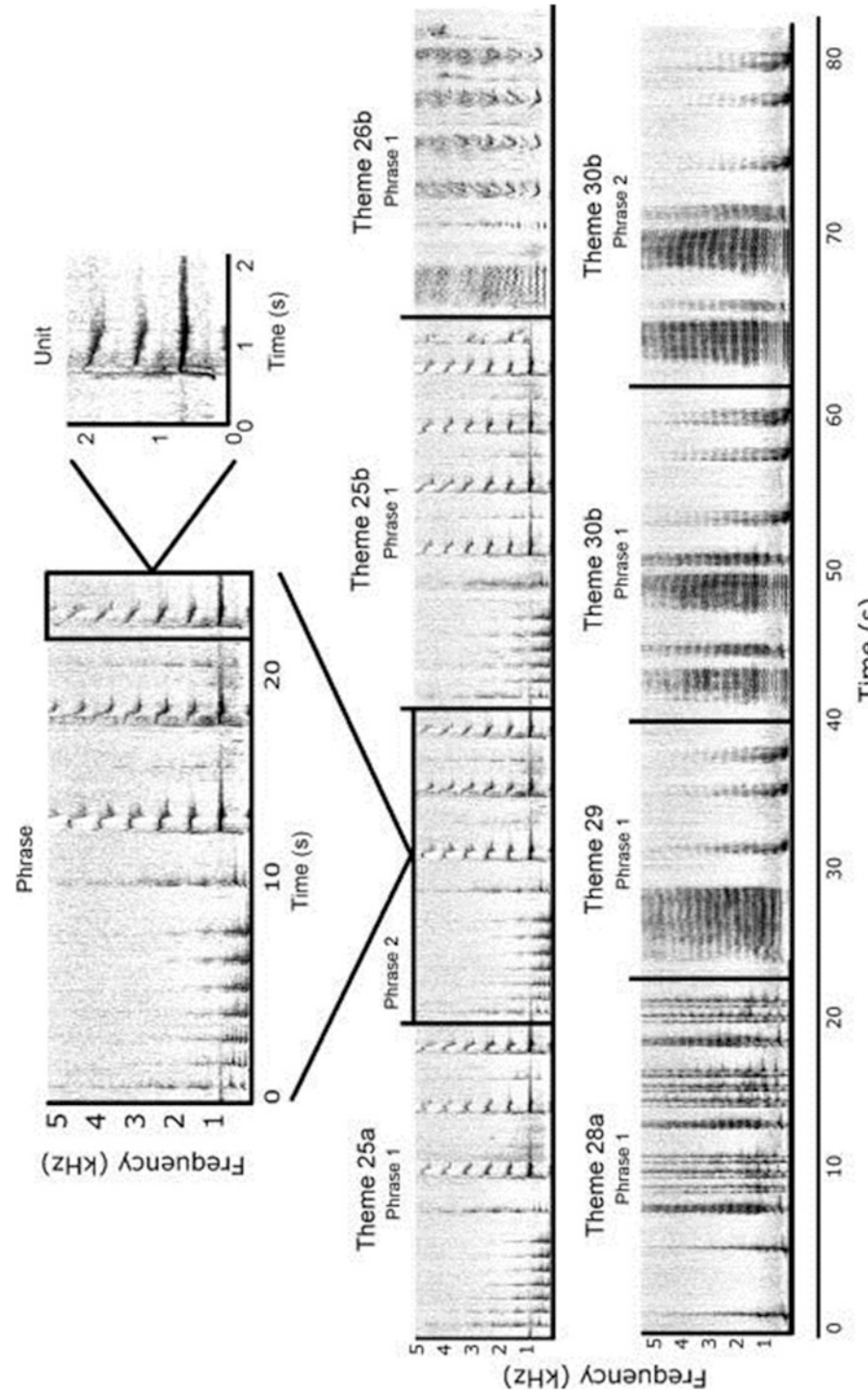


**Fig. 8.5** Spectrogram of the song structure of humpback whales, with sounds organized by theme, phrases, and units (Garland et al. 2017). © Acoustical Society of America, 2017. All rights reserved

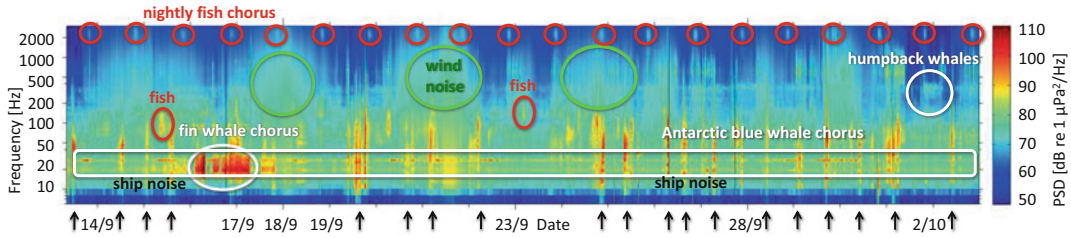


**Fig. 8.6** Spectrogram showing three weeks of choruses by fish, fin whales, and blue whales in the Perth Canyon, Australia (modified from Erbe et al. 2015). Fish raised ambient levels by 20 dB in the 1800–2500 Hz band every night. Fin whales raised ambient levels by 20 dB in the 15–35 Hz band over two days. Antarctic blue whales were the cause of ongoing tones at 18 and 28 Hz for weeks at a time. Colors represent power spectral density (PSD). Black arrows point to strong noise from passing ships. 

energy, or amplitude, of the incoming signal in a specified frequency band and to determine whether it exceeds a user-defined threshold. If the threshold within the frequency band is exceeded, the sound is scored as being present. The threshold value typically is set relative to the ambient noise in the frequency band of interest (e.g., Mellinger 2008; Ou et al. 2012). A simple energy threshold detector does not perform well when signals have low signal-to-noise ratio (SNR) or when sounds overlap. A number of techniques have been devised to overcome these problems, including spectrogram equalization (e.g., Esfahanian et al. 2017) to reduce background noise, time-varying (adaptive) detection thresholds (e.g., Morrissey et al. 2006), and using concurrent, but different, detection thresholds for different frequency bands (e.g., Brandes 2008; Ward et al. 2008). Apart from finding individual animal sounds, energy threshold detectors also have been successfully applied to the detection of animal choruses, such as those produced by spawning fish, migrating whales (Erbe et al. 2015), and chorusing insects or amphibians. These choruses are composed of many sounds from large and often distant groups of animals and so individual signals often are not detectable in them. Choruses can last for hours and significantly raise ambient levels in a species-specific frequency band (Fig. 8.6).

### 8.3.2 Spectrogram Cross-Correlation

Spectrogram cross-correlation is a well-known technique to detect sounds produced by many species, such as rockfish (genus *Sebastes*; Širović et al. 2009), African elephants (*Loxodonta africana*; Venter and Hanekom 2010), maned wolves (*Chrysocyon brachyurus*; Rocha et al. 2015), minke whales (Oswald et al. 2011), and sei whales (*Balaenoptera borealis*; Baumgartner and Fratantoni 2008). In this method, spectrograms of reference sounds from the species of interest are converted into reference coefficients, or kernels, with one kernel for each sound type (Fig. 8.7). These reference kernels then are cross-correlated with the incoming spectrogram on a frame-by-frame basis. Kernels can be a statistical representation of spectrograms of known sound types, or they can be created empirically by trial-and-error from previously analyzed recordings.

Proper selection of reference signals is critical to the performance of the detector and thus this method is only suited for detection of stereotypical sounds. Seasonal and annual variability in call structure can significantly impact performance of these detectors and so an analysis of the variability in call structure is vital when applying spectrogram cross-correlation to detect calls in long-term datasets (Širović 2016). Another

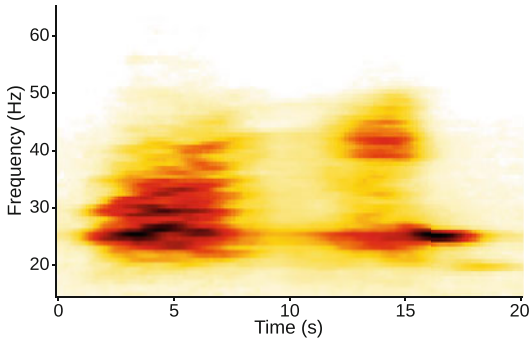


**Fig. 8.7** Spectrogram of the kernel for Omura's whales' (*Balaenoptera omurai*) doublet calls, computed as the average of over 800 hand-picked calls (Madhusudhana et al. 2020)

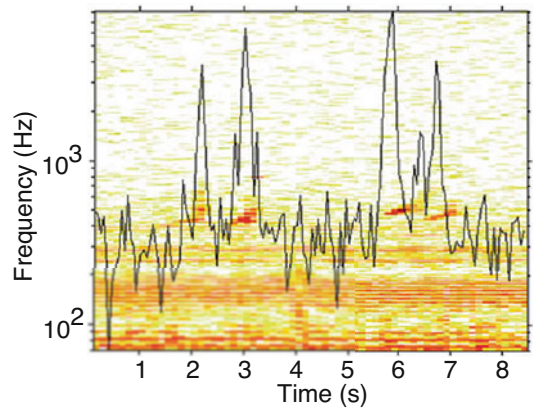


**Fig. 8.8** Spectrogram of marine mammal tonal sounds with negative entropy (black curve) overlain. Negative entropy is high when the power spectral density is concentrated in a few narrow frequency bands (Erbe and King 2008)

drawback to this method is that it can be prohibitively processor-intensive. To speed up the calculations, Harland (2008) first employed an energy threshold detector (as described above) to detect times of potential signal presence and then used spectrogram cross-correlation to detect individual signals within the flagged time periods.

### 8.3.3 Matched Filter

The matched filter approach for sound classification is similar to spectrogram cross-correlation but is performed in the time-domain. This means that the waveforms (i.e., sound pressure levels as a function of time) are correlated instead of the spectrogram. A kernel of the waveform of the sound to be detected is produced, often empirically using a high-quality recording, and then cross-correlated with the incoming signal (i.e., the time series of sound pressure). Matched filters are efficient at detecting signals in Gaussian noise (white noise), but colored noise (typical in many natural environments) poses more of a problem. As with spectrogram cross-correlation, the selection of kernels is critical to the performance of the detector. Matched filters are only appropriate for detection of well-known, stereotyped acoustic features, such as sounds produced by cane toads (Dang et al. 2008), blue whales (Stafford et al. 1998; Bouffaut et al. 2018), and beaked whales (Hamilton and Cleary 2010). Their performance suffers in the presence of even a small amount of sound variation compared to the kernel.

### 8.3.4 Spectral Entropy Detector

In general, entropy measures the disorder or uncertainty of a system. Applied to communication theory, the information entropy (also called Shannon entropy; Shannon and Weaver 1998) measures the amount of information contained in a data stream. Entropy is computed as the negative product of a probability distribution and its logarithm. Therefore, a strongly peaked probability distribution has low entropy, while a broad probability distribution has high entropy. If applied to an acoustic power spectral density distribution, entropy measures the peakedness of the power spectra and detects narrowband signals in broadband noise (Fig. 8.8). Spectral entropy has successfully been applied to animal sounds; for example, from birds, beluga whales (*Delphinapterus leucas*), bowhead whales, and walruses (Erbe and King 2008; Mellinger and Bradbury 2007; Valente et al. 2007).

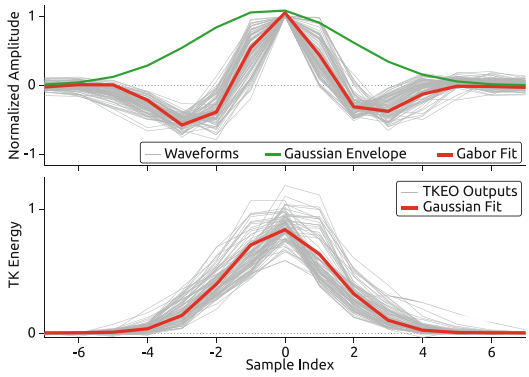


**Fig. 8.9** Waveforms of odontocete clicks and their Gabor fit (top) and TKEO outputs and Gaussian fit (bottom) (Madhusudhana et al. 2015)

### 8.3.5 Teager–Kaiser Energy Operator

The Teager–Kaiser energy operator (TKEO) is a nonlinear operator that tracks the energy of a data stream (Fig. 8.9). Operating on a time series, at any one instance, the TKEO computes the square of the sample and subtracts the product of the previous and next sample. The output is therefore high for very brief signals. The TKEO has successfully been applied to the detection of clicks, such as bat or odontocete biosonar sounds (Kandia and Stylianou 2006; Klinck and Mellinger 2011). Many biosonar signals are of Gabor type (i.e., a sinusoid modulated by a Gaussian envelope). The TKEO output of the signals is a simple Gaussian, which can be detected with simple tools, such as energy threshold detection or matched filtering (Madhusudhana et al. 2015).

### 8.3.6 Evaluating the Performance of Automated Detectors

Automated detectors can produce two types of errors: missed detections (i.e., missing a sound that exists) and false alarms (i.e., incorrectly reporting a sound that does not exist or reporting a sound that is not the target signal). There is an inevitable trade-off when choosing the acceptable rate of each. Most detectors allow the user to adjust a threshold, and depending on where this threshold is set, the probability of one type of error increases while the other decreases. The acceptability of either type of error is determined by the particular application of the detector. For example, for rare animals in critical habitats, detecting every sound, even those that are very faint, is desired. In this situation, a low threshold can be chosen that minimizes the number of missed detections; however, this can result in many false alarms. Quantification of these two errors is a useful way to evaluate the performance of an automated detector.

| | | Detector Input | |
|---|---|---|---|
| | | Signal Present | Signal Absent |
| Reported Output | Signal Present | True Positive (TP) Correct Detection | False Positive (FP) False Alarm |
| | Signal Absent | False Negative (FN) Missed Detection | True Negative (TN) Correct Rejection |

**Fig. 8.10** Confusion matrix showing the possible outcomes of a detector when a signal is present versus absent

#### 8.3.6.1 Confusion Matrices

One of the simplest and most common methods for conveying the performance of a detector (or a classifier) is a confusion matrix (i.e., a type of contingency table). A confusion matrix (Fig. 8.10) gives the number of true positives (i.e., correctly classified sounds, also called correct detections), false positives (i.e., false alarms), true negatives (i.e., correct rejections), and false negatives (i.e., missed detections).

#### 8.3.6.2 Receiver Operating Characteristic (ROC) Curve

The performance of detectors can be visualized using the receiver operating characteristic (ROC) curve. A ROC curve is a graph that depicts the trade-offs between true positives and false positives (Egan 1975; Swets et al. 2000). The false positive rate (i.e., FP/(FP+TN)) is plotted on the x-axis, while the true positive rate (i.e., TP/(TP +FN)) is plotted on the y-axis (Fig. 8.11). A curve is generated by plotting these values for the detector at different threshold values. The (0l1) point on the graph represents perfect performance: 100% true positives and no false positives.

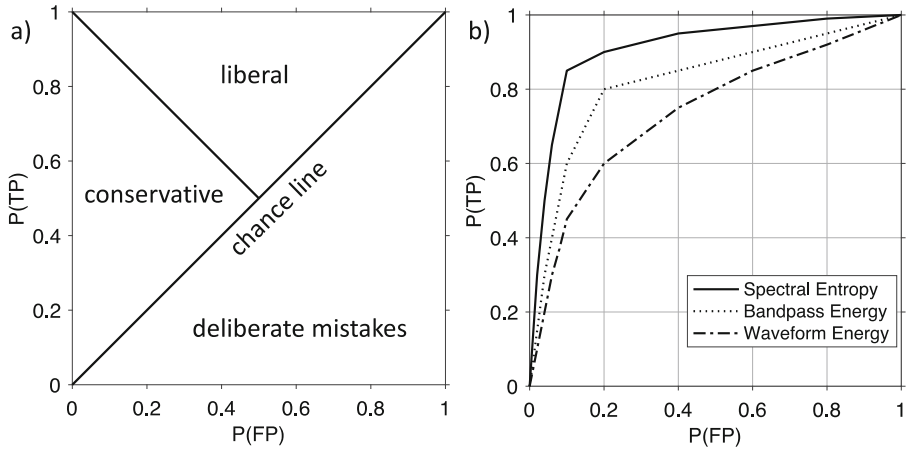


**Fig. 8.11** (**a**) Generalized receiver operating characteristic (ROC) plot, in which the probability of true positives is plotted against the probability of false positives. Areas in this graph that correspond to a liberal bias, conservative bias, and deliberate mistakes are indicated. (**b**) Example ROC curves computed during the development of automated detectors for marine mammal calls in the Arctic. The spectral entropy detector outperformed others (Erbe and King 2008)

The major diagonal in Fig. 8.11a represents performance at chance, where the probabilities of TP and FP are equal. Responses falling below the line would indicate deliberate mistakes. The minor diagonal represents neutral bias, and splits responses into conservative versus liberal. A conservative response strategy yields decreased correct detection and false alarm probabilities; a liberal response strategy yields increased correct detection and false alarm probabilities. An example ROC curve is given in Fig. 8.11b, comparing the performances of three detectors (operating on underwater acoustic recordings from the Arctic and trying to detect marine mammal calls) based on: (1) spectral entropy, (2) bandpassed energy, and (3) waveform (i.e., broadband) energy. The performance of the entropy detector surpassed that of the other two.

#### 8.3.6.3 Precision and Recall

The performance of a detector can be overestimated using a ROC curve when there is a large difference between the numbers of TPs and TNs. In addition, estimation of the number of TNs requires discrete sampling units. The duration of the discrete sampling units is often somewhat arbitrary and can lead to unrealistic differences between the numbers of TPs and TNs. In these situations, precision and recall (P-R) can provide a more accurate representation of detector performance because this representation does not rely on determining the number of true negatives (Davis and Goadrich 2006). In the P-R framework, events are scored only as TPs, FPs, and FNs.

Precision is a measure of accuracy and is the proportion of automated detections that are true detections.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is a measure of completeness and is the proportion of true events that are detected. This is the same as the true positive rate defined in the ROC framework.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Detectors can be evaluated by plotting precision against recall (Fig. 8.12). An ideal detector would have both scores approaching a value of 1. In other words, the curve would approach the upper right-hand corner of the graph (Davis and Goadrich 2006). Precision and recall also can be

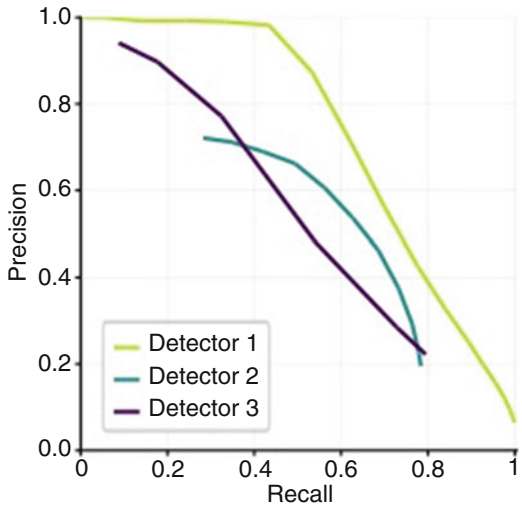


**Fig. 8.12** Precision-Recall curves for three types of detectors: (1) spectrogram cross-correlation, (2) blob detection, and (3) spectral entropy for Omura's whale calls (Madhusudhana et al. 2020)

represented by an *F*-score, which is the geometric mean of these values. The F-score can be weighted to emphasize either precision or recall when optimizing detector performance (Jacobson et al. 2013).

## 8.4 Quantitative Classification of Animal Sounds

Quantitative classification of animal sounds is based on measured features of sounds, no matter whether these are used to manually or automatically group sounds with the aid of software algorithms. These features can be measured from different representations of sounds, such as waveforms, power spectra, spectrograms, and others. A large variety of classification methods have been applied to animal sounds, many drawing from human speech analysis.

### 8.4.1 Feature Selection

The acoustic features selected and the consistency with which the measurements are taken have a significant influence on the success (or failure) of a classification algorithm. Feature sets (also called feature vectors) should provide as much information as sensible about the sounds. With today's software tools and computing power, a limitless number of features can easily be measured that would allow distinction between sounds even of the same type. Such over-parameterization can make it difficult to group like sounds, which can be just as important as distinguishing between different sounds. The challenge is to find the trade-off and produce a set of representative features for each sound type. Once the features have been selected, automating the extraction and subsequent analysis of these features reduces the time required to analyze large datasets. Some commonly used feature vectors are described below.

#### 8.4.1.1 Spectrographic Features

Perhaps the most commonly used feature vectors are those consisting of values measured from spectrograms. These measurements include, but are not limited to, frequency variables (e.g., frequency at the beginning of the sound, frequency at the end of the sound, minimum frequency, maximum frequency, frequency of peak energy, bandwidth, and presence/absence of harmonics or sidebands; Fig. 8.13; also see Chap. 4, Sect. 4.2.3), and time variables (e.g., signal duration, phrase and song length, inter-signal intervals, and repetition rate). More complex features, such as those describing the spectrographic shape of a sound (e.g., upsweep, downsweep, chevron, U-loop, inverted U-loop, or warble), slopes, and numbers and relative positions of local extrema and inflection points (places where the contour changes from positive to negative slope or vice versa) also have been used in classification. These measurements often are taken manually from spectrographic displays (e.g., by a technician using a mouse-controlled cursor). Automated techniques for extracting spectrographic measurements can be less subjective and less time-consuming, but are sometimes not as accurate as manual methods. Examples are available in the bird literature (e.g., Tchernichovski et al. 2000), bat literature (Gannon et al. 2004; O'Farrell et al. 1999), and marine mammal

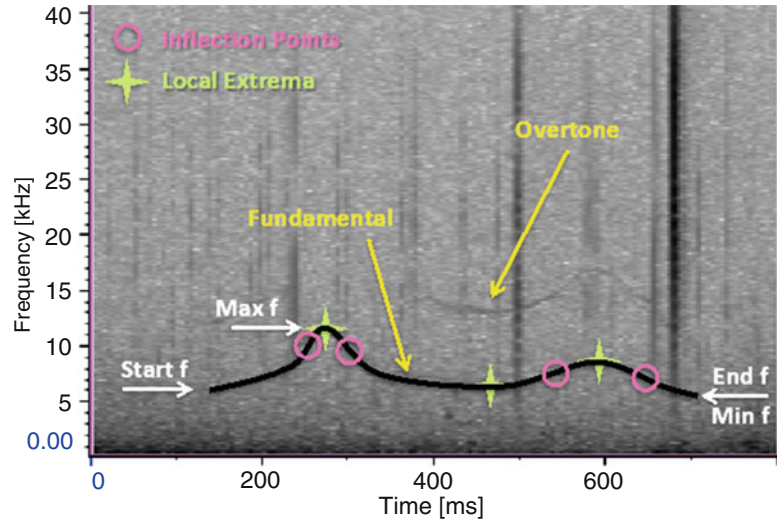


**Fig. 8.13** Spectrogram of a pilot whale (*Globicephala melas*) whistle showing the following features: Start frequency (Start f), End frequency (End f), Maximum frequency (Max f), Minimum frequency (Min f), locations of two local maxima and one local minimum in the fundamental contour, four inflection points (where the curvature changes from clockwise to counter-clockwise, or vice versa), and one overtone (Courts et al. 2020). © Courts et al.; https://www.nature.com/articles/s41598-020-74111-y/figures/5. Licensed under CC BY 4.0; https://creativecommons.org/licenses/by/4.0/

literature (e.g., Mellinger et al. 2011; Roch et al. 2011; Gillespie et al. 2013; Kershenbaum et al. 2016). Spectrographic measurements of bat calls, for example, can be extracted using *Analook* (Titley Scientific, Columbia, MO, USA), *SonoBat* (Joe Szewczak, Department of Biology, Humboldt State University, Arcata, CA, USA), or *Kaleidoscope* Pro (Wildlife Acoustics, Inc., Maynard, MA, USA), exported to an *Excel* spreadsheet (XML, CSV, and other formats), classified using machine learning algorithms, and compared to a reference library for identification.

#### 8.4.1.2 Cepstral Features

Cepstral coefficients are spectral features of bioacoustic signals commonly used in human speech processing (Davis and Mermelstein 1980). These features are based on the source-filter model of human speech analysis, which has been applied to many different animal species (Fitch 2003). Cepstral coefficients are well-suited for statistical pattern-recognition models because they tend to be uncorrelated (Clemins et al. 2005), which significantly reduces the number of parameters that must be estimated (Picone 1993). Cepstral coefficients are calculated by computing the Fourier transform in successive time windows over the recorded pressure time series of a sound (see Chap. 4). The frequency axis then is warped by multiplying the spectrum with a series of $n$ filter functions at appropriately spaced frequencies. This is done because there is evidence that many animals perceive frequencies on a logarithmic scale, in a similar fashion to humans (Clemins et al. 2005). The output of the frequency band filters is then used as input to a discrete cosine transform, which results in an $n$-dimensional cepstral feature vector (Picone 1993; Clemins et al. 2005; Roch et al. 2007, 2008).

Using cepstral feature space allows the timbre of sounds to be captured, a quality that is lost when extracting parameters from spectrograms. Roch et al. (2007) developed an automated classification system based on cepstral feature vectors extracted for whistles, burst-pulse sounds, and clicks produced by short- and long-beaked

common dolphins (*Delphinus* spp.), Pacific white-sided dolphins (*Lagenorhynchus obliquidens*), and bottlenose dolphins (*Tursiops truncatus*). The system did not rely on specific sound types and had no requirement for separating individual sounds. The system performed relatively well, with correct classification scores of 65–75%, depending on the partitioning of the training- and test-data. Cepstral feature vectors also have been used as input to classifiers for many other animal species, including groupers (*Epinephelus guttatus, E. striatus, Mycteroperca venenosa, M. bonaci;* Ibrahim et al. 2018), frogs (Gingras and Fitch 2013), song birds (Somervuo et al. 2006), African elephants (Zeppelzauer et al. 2015), and beluga, bowhead, gray (*Eschrichtius robustus*), humpback, and killer (*Orcinus orca*) whales, and walrus (Mouy et al. 2008). Cepstral features appear to be a promising alternative to the traditional time- and frequency-parameters measured from spectrograms as input to classification algorithms. However, cepstral features are relatively sensitive to the SNR, the signal's phase, and modeling order (Ghosh et al. 1992).

Noda et al. (2016) used mel-frequency cepstral coefficients and random forest analyses to classify sounds produced by 102 species of fish and compared the performance of three classifiers: k-nearest neighbors, random forest, and support vector machines (SVMs). The mel-frequency cepstrum (or cepstrogram) is a form of acoustic power spectrum (or spectrogram) that is computed as a linear cosine transform of a log-power spectrum that is presented on a nonlinear mel-scale of frequency. The mel-scale resembles the human auditory system better than the linearly-spaced frequency bands of the normal cepstrum. All three classifiers performed similarly, with average classification accuracy ranging between 93% and 95%.

### 8.4.2 Statistical Classification of Animal Sounds

For some sounds, qualitative classification is sufficient. Janik (1999) reported that humans were able to identify dolphin signature whistles more reliably than computer methods. A problem with qualitative classification of sounds in a repertoire (and taxonomy in general), however, is that some listeners are "splitters" and other listeners are "lumpers." So, even researchers on the same project could classify an animal's sound repertoire differently. One way to avoid individual researcher differences in classification is to use graphical, statistical, and computer-automated methods that objectively sort and compare measured variables that describe the sounds. A variety of statistical methods can be employed to classify animal sounds into categories (Frommolt et al. 2007). Below are brief descriptions of some of the statistical methods that are commonly used for classification of animal sounds.

#### 8.4.2.1 Parametric Clustering

Parametric cluster analysis produces a dendrogram (i.e., classification tree) that organizes similar sounds into branches of a tree. A distance matrix also is generated, which gives correlation coefficients between all variables in the dataset. The resulting distance index ranges from 0 (very similar sounds) to 1 (totally dissimilar sounds). The matrix can then be joined by rows or columns to examine relationships. The type of linkage and type of distance measurement can be selected to find the best fit for a particular dataset (Zar 2009).

Cluster analysis has been used to classify sound types in several species, including owls (Nagy and Rockwell 2012), mice (Hammerschmidt et al. 2012), rats (*Rattus norvegicus*, Takahashi et al. 2010), African elephants (Wood et al. 2005), and primates (Hammerschmidt and Fischer 1998). In a study of six populations of the neotropical frog (*Proceratophrys moratoi*) in Brazil, Forti et al. (2016) measured spectrographic variables from calls produced by males and performed cluster analysis to examine similarities in acoustic traits (based on the Bray–Curtis index of acoustic similarity) across the six locations (Fig. 8.14). Baptista and Gaunt (1997) used hierarchical cluster analysis of correlation coefficients of several acoustic parameters to categorize sounds of the sparkling violet-eared hummingbird (*Colibri*

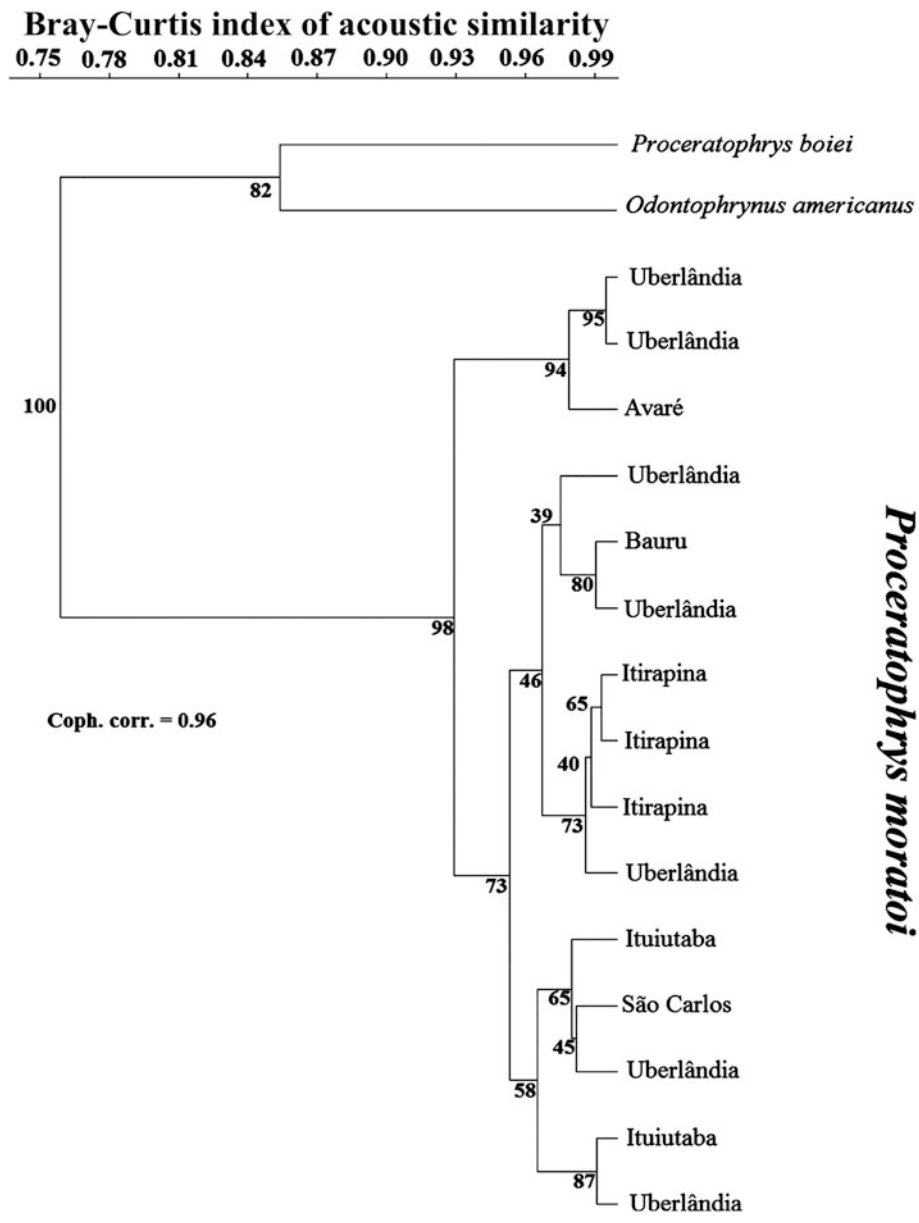


**Fig. 8.14** Dendrogram from a hierarchical cluster analysis of the call similarities between 15 male *Proceratophrys moratoi* from different sites and two other Odontophrynidae species (Forti et al. 2016). © Forti et al.; https://peerj.com/articles/2014/. Licensed under CC BY 4.0; https://creativecommons.org/licenses/by/4.0/

*coruscans*), which is found in two neighboring assemblages in their study area. A matrix of sound similarity values obtained from spectral cross-correlation of these birds' songs indicated similar sound types from the two areas. Yang et al. (2007) used cluster analysis to examine syllable sharing between individuals of Anna's hummingbird (*Calypte anna*). They identified 38 syllable types in songs of 44 males, which clustered into five basic syllable categories: "Bzz," "bzz," "chur," "ZWEE," and "dz!". Also, microgeographic song variation patterns were

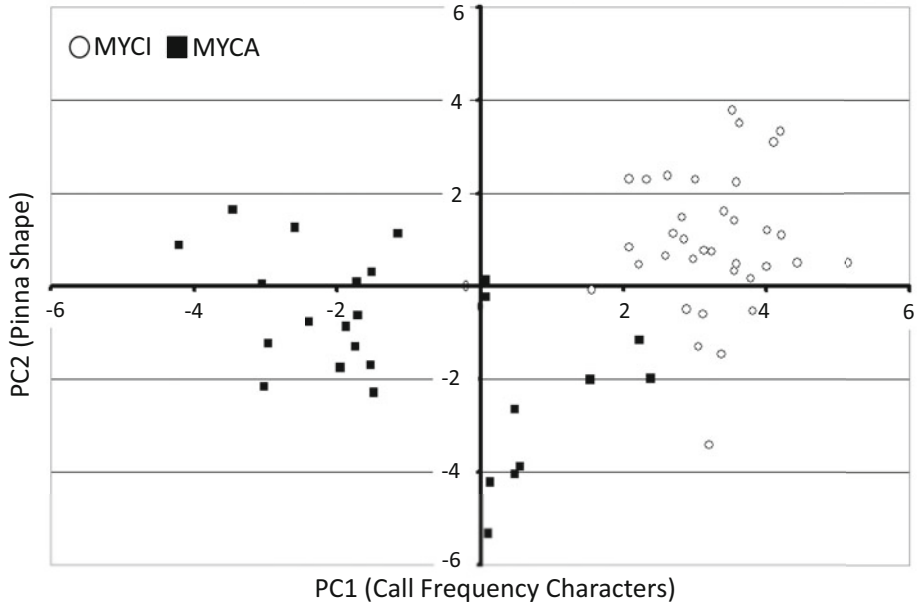


**Fig. 8.15** Plot showing the results of principal component analysis, in which two cryptic species of myotis bats (California myotis, *Myotis californicus,* MYCA, black squares; western small-footed bat, *M. ciliolabrum,* MYCI, hollow circles) were distinguished by differences in ear height and characteristic frequency of their echolocation signals. Plotted is characteristic frequency versus signal duration for these species recorded from field sites in New Mexico and Arizona, USA

found in that nearest neighbors sang more similar songs than non-neighbors. Pozzi et al. (2010) used several acoustic variables to group black lemur (*Eulemur macaco macaco*) sounds into categories, including the frequencies of the fundamental and of the first three harmonic overtones (measured at the start, middle, and end of each call), and the total duration. The agreement of this analysis with manual classification was high (>88.4%) for six of eight categories.

#### 8.4.2.2 Principal Component Analysis

Principal component analysis (PCA) is a multivariate statistical method that examines a set of measurements such as the feature vectors discussed earlier in Sect. 8.4. These features may well be correlated. For example, bandwidth is sometimes correlated with maximum frequency, or the number of inflection points can be correlated with signal duration (Ward et al. 2016). PCA performs an orthogonal transformation that converts the potentially correlated variables (i.e., the features) into a set of linearly uncorrelated variables (i.e., the principal components; Hotelling 1933; Zar 2009). The principal components are linear combinations of the original variables (features). Plotting the principal components against each other shows how the measurements cluster.

For example, by examining bat biosonar signals in multivariate space, bat species that are very similar in external appearance can be distinguished. Using PCA, Gannon et al. (2001) found ear height and characteristic frequency were correlated, along with duration of the signal (Fig. 8.15).

As another example, Briefer et al. (2015) categorized emotional states associated with variation in whinnies from 20 domestic horses (*Equus ferus*) using PCA. They designed four situations to elicit different levels of emotional arousal that were likely to stimulate whinnies: separation (negative situation) and reunion (positive situation) with either all group members (high

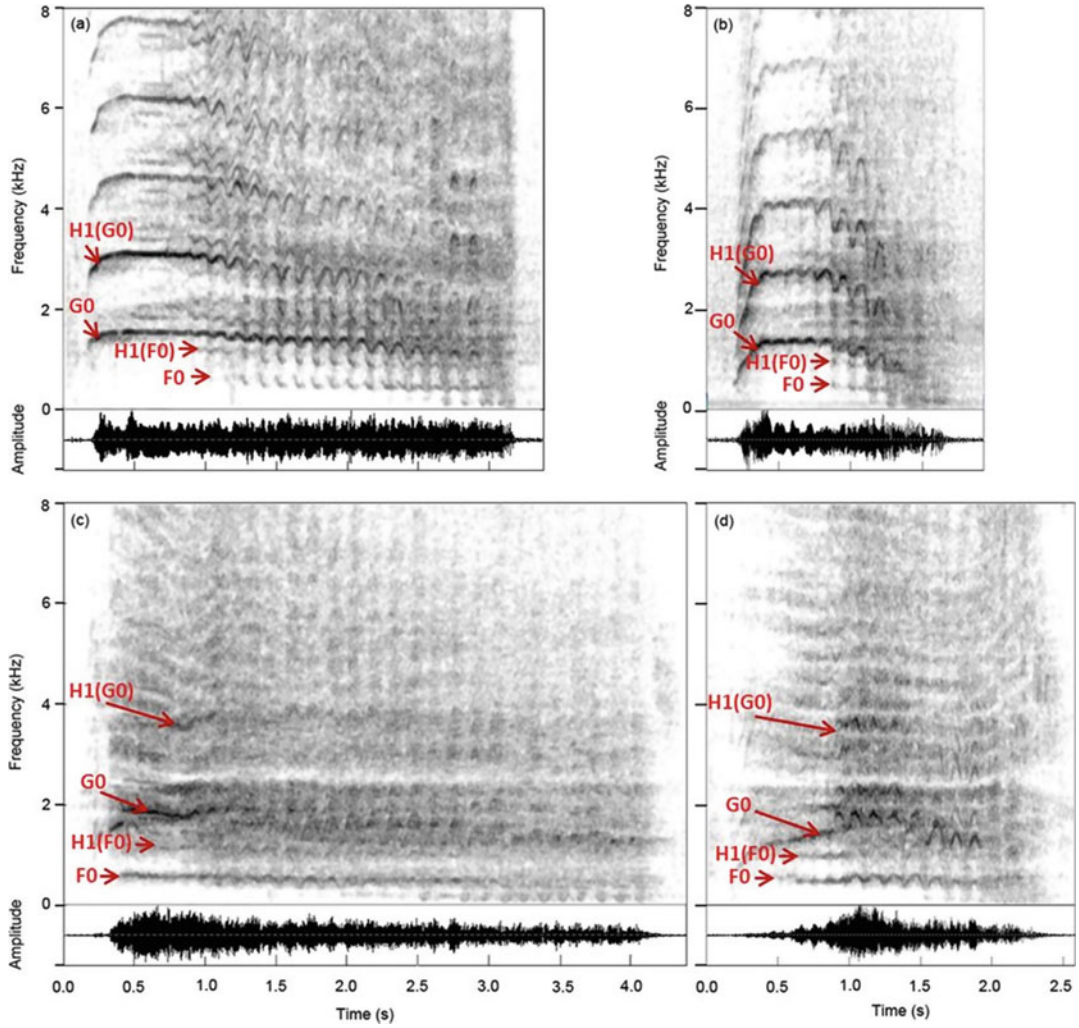


**Fig. 8.16** Spectrograms and oscillograms of horse whinnies in negative (**a**, **c**) and positive (**b**, **d**) situations emitted by two different horses. Red arrows point to fundamental frequencies (F0, G0) and first overtones (H1). Negative whinnies (**a**, **c**) are longer in duration and have higher G0 fundamentals than positive whinnies (**b**, **d** Briefer et al. 2015). © Briefer et al.; https://www.nature.com/articles/srep09989/figures/3. Licensed under CC BY 4.0; http://creativecommons.org/licenses/by/4.0/

emotional arousal) or only one group member (moderate emotional arousal). The authors measured 21 acoustic features from whinnies (Fig. 8.16). PCA transformed the feature vectors into six principal components that accounted for 83% of the variance in the original dataset.

#### 8.4.2.3 Discriminant Function Analysis

In discriminant function analysis (DFA), canonical discriminant functions are calculated using variables measured from a training dataset. One canonical discriminant function is produced for each sound type in the dataset. Variables measured from sounds in the test dataset are then substituted into each function and each sound type is classified according to the function that produced the highest value. Because DFA is a parametric technique, it is assumed that input data have a multivariate normal distribution with the same covariance matrix (Afifi and Clark 1996;

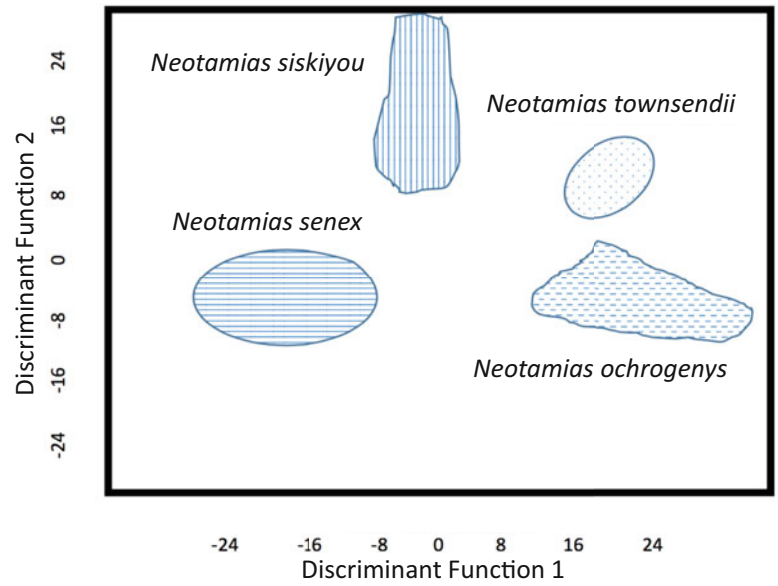


**Fig. 8.17** Plot resulting from discriminant function analysis. Four species of Townsend-group chipmunks (Townsend's chipmunk, *Neotamias townsendii*; Siskiyou chipmunk, *N. siskiyou*; Allen's chipmunk, *N. senex*; and yellow-cheeked chipmunk, *N. ochrogenys*) in northern California, USA, produced discernibly different sounds. Discriminant function 1 was dominated by differences in maximum frequency of the signal and discriminant function 2 was most influenced by temporal features including total signal length and the number of signals emitted by a chipmunk during a signaling bout

Zar 2009). Violations of these assumptions can create problems with some datasets. One of the main weaknesses of DFA for animal sound classification is that it assumes classes are linearly separable. Because a linear combination of variables takes place in this analysis, the feature space can only be separated in certain, restricted ways that are not appropriate for all animal sounds. Figure 8.17 shows the DFA separation of California chipmunk (genus *Neotamias*) taxa that are morphologically similar but acoustically different, using six variables measured from their sounds.

#### 8.4.2.4 Classification Trees

Classification tree analysis is a non-parametric statistical technique that recursively partitions data into groups known as "nodes" through a series of binary splits of the dataset (Clark and Pregibon 1992; Breiman et al. 1984). Each split is based on a value for a single variable and the criteria for making splits are known as primary splitting rules. The goal for each split is to divide the data into two nodes, each as homogeneous as possible. As the tree is grown, results are split into successively purer nodes. This continues until each node contains perfectly homogeneous data (Gillespie and Caillat 2008). Once this maximal tree has been generated, it is pruned by removing nodes and examining the error rates of these smaller trees. The smallest tree with the highest predictive accuracy is the optimal tree (Oswald et al. 2003).

Tree-based analysis provides several advantages over some of the other classification techniques. It is a non-parametric technique; therefore, data do not need to be normally distributed as required for other methods, such as DFA. In addition, tree-based analysis is a simple and naturally intuitive way for humans to classify sounds. It is essentially a series of true/false questions, which makes the classification process transparent. This allows easy examination of which variables are most important in the classification process. Tree-based analysis also

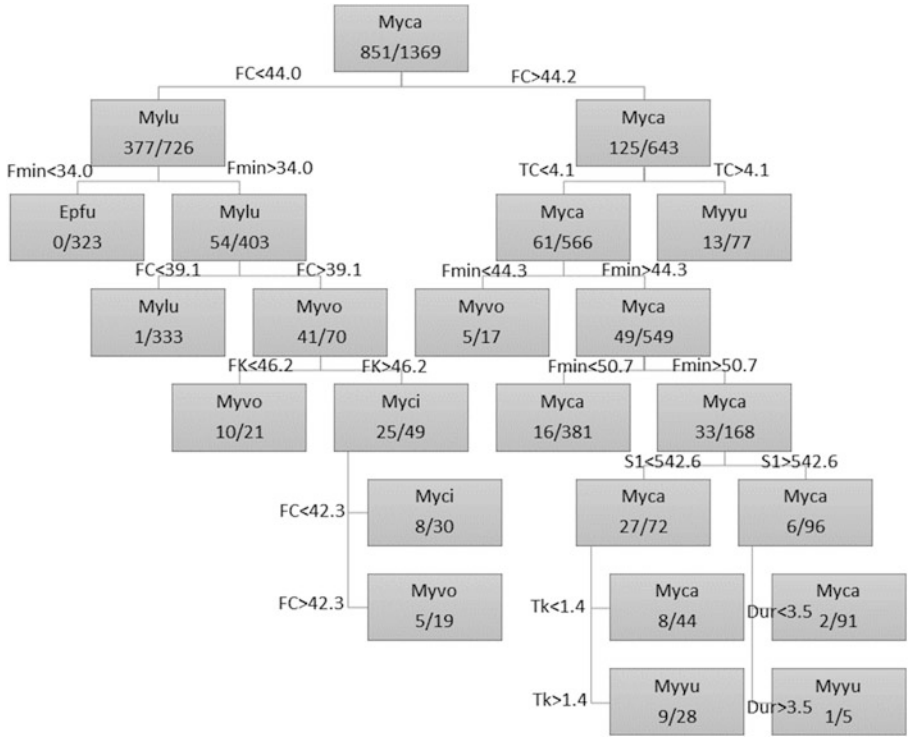


**Fig. 8.18** Classification tree grown using *Splus* computer software (version S-PLUS 6.2 2003, TIBCO Software Inc., Palo Alto, CA, USA) from 1369 bat calls. The pruned tree used variables measured from each bat call: duration (DUR), minimum frequency (Fmin), characteristic frequency (Fc; i.e., frequency at the flattest part of the call), frequency at the "knee" of the call (Fk), time of Fc, time at Fk, and slope (S1). Along the tangents between boxes are values for variables used to split the nodes (for instance, Fmin is minimum frequency). The fraction below each box is the misclassification rate (e.g., 1/5 = 20% misclassification rate). The tree has 12 terminal nodes defining the branches, resulting in a classification designation for each species (Gannon et al. 2004)

accommodates for a high degree of diversity within classes. For example, if a species produces two or more distinct sound types, a tree-based analysis can create two different nodes. In other classification techniques, different sound types within a species simply act to increase variability and make classification more difficult. Finally, surrogate splitters are provided at each node (Oswald et al. 2003). Surrogate splitters closely follow primary splitting rules and can be used in cases when the primary splitting variable is missing. Therefore, sounds can be classified even if data for some variables are missing due to noise or other factors.

To address some controversy as to whether closely related species of myotis bats could be differentiated by their sounds, Gannon et al. (2004) completed an analysis of echolocation pulses from free-flying, wild bats. Fig. 8.18 is a classification tree grown from nearly 1400 calls using at least seven variables measured from each call. The tree produced terminal nodes identified to species (MYVO is *Myotis volans*, MYCA *M. californicus,* etc.). In this study, recordings were made under field conditions where sounds were affected by the environment, Doppler shift, and diversity of equipment. Still, classification trees worked well to predict group membership and additional techniques, such as DFA, were able to distinguish five *Myotis* species acoustically with greater than 75% accuracy (greater than 90% in most instances).

Classification trees have been applied to marine mammal sounds by several researchers,

with promising results. Fristrup and Watkins (1993) used tree-based analysis to classify the sounds of 53 species of marine mammal (including mysticetes, odontocetes, pinnipeds, and manatees). Their correct classification score of 66% was 16% higher than the score obtained when applying DFA to the same dataset. The whistles of nine delphinid species were correctly classified 53% of the time by Oswald et al. (2003) using tree-based analysis. Oswald et al. (2007) subsequently applied classification tree analysis to the whistles of seven species and one genus of marine mammal, resulting in a correct classification score of 41%. This score was improved slightly, to 46%, when classification decisions were based on a combination of classification tree and DFA results. Gannier et al. (2010) used classification trees to identify the whistles of five delphinid species recorded in the Mediterranean, with a correct classification score of 63%. Finally, Gillespie and Caillat (2008) classified the clicks of Blainville's beaked whales (*Mesoplodon densirostris)*, short-finned pilot whales (*Globicephala macrorhynchus*), and Risso's dolphins (*Grampus griseus*). Their tree-based analysis classified 80% of clicks to the correct species.

#### 8.4.2.5 Nonlinear Dimensionality Reduction

Clustering techniques described above require that certain features or measurements, as appropriate for the problem domain, be available beforehand. They are gathered from sound recordings either manually (e.g., number of inflection points in whistle contours, number of harmonics) or using signal processing tools (e.g., peak frequency, energy), or both. Manual extraction of features is usually time-consuming and often inefficient, especially when dealing with recordings covering large spatial and temporal scales. Automated extraction of measurements improves efficiency and eliminates the risk of human biases. However, when recordings contain a lot of confounding sounds or have extreme noise variations, reliability and accuracy of the measurements can become questionable and can have adverse effects on clustering outcomes. Regardless of whether manual or automated approaches were employed, the resulting limited set of chosen features or measurements are essentially representations of the underlying data in a reduced space. Such dimensionality reduction is typically aimed at making the downstream task of clustering (with PCA, DFA, etc.) computationally tractable.

In recent years, nonlinear dimensionality reduction methods have gained widespread popularity, specifically in applications for exploring and visualizing very high-dimensional data. Originally popular for processing image-like data in the field of machine learning, these methods bring about dimensionality reduction without requiring one to explicitly choose and extract features. The methods can be easily adapted for processing bioacoustic recordings wherein the qualitative cluster structure (i.e., similarities in the visually identifiable information) in spectrogram-like data (e.g., mel-spectrogram or cepstrogram) containing hundreds or thousands of time-frequency points is effectively captured in an equivalent 2- or 3-dimensional space (e.g., Sainburg et al. 2019; Kollmorgen et al. 2020).

One of the earlier methods for capturing nonlinear structure, the t-distributed stochastic neighbor embedding (t-SNE; van der Maaten and Hinton 2008) is based on non-convex optimization. It computes a similarity measure between pairs of points (data samples) in the original high-dimensional space and in the reduced space, then minimizes the Kullback–Leibler divergence between the two sets of similarity measures. t-SNE tries to preserve distances in a neighborhood whereby points close together in the high-dimensional space have a high probability of staying close in the reduced space. The *Bird Sounds* project (Tan and McDonald 2017) presents an excellent demonstration of using t-SNE for organizing thousands of bird sound spectrograms in a 2-dimensional similarity grid.

Some of the shortcomings of t-SNE were addressed in a newer method called uniform manifold approximation and projection (UMAP; McInnes et al. 2018). UMAP is backed with a strong theoretical framework. While effectively capturing local structures like t-SNE, UMAP

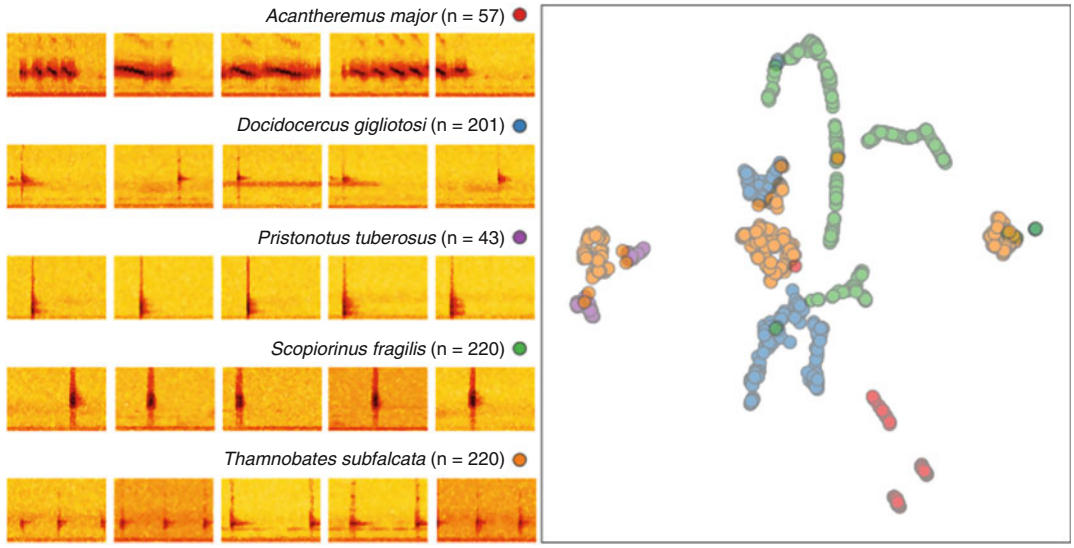


**Fig. 8.19** Demonstration of clustering katydid sounds using UMAP. Randomly chosen samples of call spectrograms of the five species considered are shown on the left, and clustering outcomes are shown on the right. The clustering activity has successfully captured both inter-species and intra-species variations

also offers a better promise for preserving global structures (inter-cluster relationships). UMAP processes data faster and is capable of handling very large dimensional data. Fig. 8.19 is a demonstration of the use of UMAP for clustering sounds of five species of katydids (Tettigoniidae) from Panamanian rainforest recordings (Madhusudhana et al. 2019). Inputs to UMAP clustering comprised of spectrograms (dimensions 216h x 469w) computed from 1-s clips containing katydid call(s). The inputs often contained confounding sounds and varying noise levels. The clustering results, however, demonstrate the utility of UMAP as a quick means to effective clustering. UMAP has also been used, in combination with a pre-trained neural network, for assessing habitat quality and biodiversity variations from soundscape recordings across different ecosystems (Sethi et al. 2020).

We have presented here two popular methods that are currently trending in this field of research. There are, however, other alternatives available including earlier methods such as isomap (Tenenbaum et al. 2000) and diffusion map (Coifman et al. 2005), newer variants of t-SNE (e.g., Maaten 2014; Linderman et al. 2017), and some modern variants of variational autoencoders (Kingma and Welling 2013).

## 8.4.3 Model Based Classification

### 8.4.3.1 Artificial Neural Networks

Artificial neural networks (ANNs) were developed by modeling biological systems of information-processing (Rosenblatt 1958) and became very popular in the areas of word recognition in human speech studies (e.g., Waibel et al. 1989; Gemello and Mana 1991) and character or image-recognition (e.g., Fukushima and Wake 1990; Van Allen et al. 1990; Belliustin et al. 1991) in the 1980s. Since that time, ANNs have been used successfully to classify a number of complex signal types, including quail crows (*Coturnix* spp., Deregnaucourt et al. 2001), alarm sounds of Gunnison's prairie dogs (*Cynomys gunnisoni*, Placer and Slobodchikoff 2000), stress sounds by domestic pigs (*Sus scrofa domesticus*, Schon et al. 2001), and dolphin echolocation clicks (Roitblat et al. 1989; Au and Nachtigall 1995).

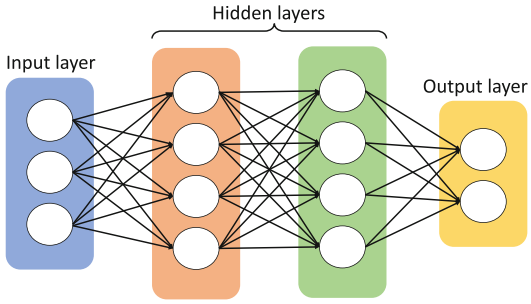


**Fig. 8.20** Diagram of the structure of an artificial neural network

In their primitive forms, there are 20 or more basic architectures of ANNs (see Lippman 1989 for a review). Each ANN approach results in trade-offs in computer memory and computation requirements, training complexity, and time and ease of implementation and adaptation (Lippman 1989). The choice of ANN depends on the type of problem to be solved, size and complexity of the dataset, and the computational resources available. All ANNs are composed of units called neurons and connections among them. They typically consist of three or more neuron layers: one input layer, one output layer, and one or more hidden layers (Fig. 8.20). The input layer consists of $n$ neurons that code for $n$ features in the feature vector representing the signal ($X_1 \ldots X_n$). The output layer consists of $k$ neurons representing the $k$ classes. The number of hidden layers between the input and output layers, as well as the number of neurons per layer, is empirically chosen by the researcher. Each connection among neurons in the network is associated with a weight-value, which is modified by successive iterations during the training of the network.

ANNs are promising for automatic signal classification for several reasons. First, the input to an ANN can range from feature vectors of measurements taken from spectrograms or waveforms, to frequency contours, to complete spectrograms. Second, ANNs serve as adaptive classifiers which learn through examples. As a result, it is not necessary to develop a good mathematical model for the underlying signal characteristics before analysis begins (Ghosh et al. 1992). In addition, ANNs are nonlinear estimators that are well-suited for problems involving arbitrary distributions and noisy input (Ghosh et al. 1992; Potter et al. 1994).

Dawson et al. (2006) used artificial neural networks as a means to classify the chick-a-dee-dee-dee call of the black-capped chickadee *(Poecile atricapillus)*, which contains four note types carrying important functional roles in this species. In their study, an ANN first was trained to identify the note type based on several acoustic variables and then correctly classified recordings of the notes with 98% accuracy. The performance of the network was compared with classification using DFA, which also achieved a high level of correct classification (95%). The authors concluded that "there is little reason to prefer one technique over another. Either method would perform extremely well as a note-classification tool in a research laboratory" (Dawson et al. 2006).

Placer and Slobodchikoff (2000) used artificial neural networks to classify alarm sounds of Gunnison's prairie dogs (*Cynomys gunnisoni*) to predator species with a classification accuracy of 78.6 to 96.3%. The ANN identified unique signals for four different species of predators: red-tailed hawk (*Buteo jamaicensis*), domestic dog (*Canis familiaris*), coyote (*Canis latrans*), and humans (*Homo sapiens*).

Deecke et al. (1999) used artificial neural networks to examine dialects in underwater sounds of killer whale pods. The neural network extracted the frequency contours of one sound type shared by nine social groups of killer whales and created a neural network similarity index. Results were compared to the sound similarity judged by three humans in pair-wise classification tasks. Similarity ratings of the neural network mostly agreed with those of the humans, and were significantly correlated with the killer whale group, indicating that the similarity indices were biologically meaningful. According to the authors, "an index based on neural network analysis therefore represents an objective and repeatable means of measuring acoustic similarity, and allows comparison of results across studies, species, and time" (Deecke et al. 1999).

The greater potential of ANNs remained largely untapped for many years, in part due to prevailing limitations in computational capabilities. In the mid-1980s, backpropagation paved a way for efficiently training multi-layer ANNs (Rumelhart et al. 1986). Backpropagation, an algorithm for supervised learning of the weights in an ANN using gradient descent, greatly facilitated development of deeper networks (having many hidden layers). Many classes of deep neural networks (DNNs; LeCun et al. 2015) such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) became easier to train. While the aforementioned ANN approaches often require hand-picked features or measurements as inputs, DNNs trained with backpropagation demonstrated the ability to learn good internal representations from raw data (i.e., the hidden layers captured non-trivial representations effectively). In their landmark work on using CNNs for the automatic recognition of handwritten digits, LeCun et al. (1989a, b) used backpropagation to learn convolutional kernel coefficients directly from images. Over the past two decades, advances in computing technology, especially the wider availability of graphics processing units (GPUs), have considerably accelerated machine learning (ML) research in many disciplines such as computer vision, speech processing, natural language processing, recommendation systems, etc. Shift invariance is an attractive characteristic of CNNs, which makes them suitable for analyzing visual imagery (LeCun et al. 1989a, b, 1998). CNN-based solutions have consistently dominated many of the large-scale visual recognition challenges. As such, several competing architectures of CNNs have been developed: AlexNet (Krizhevsky et al. 2017), ResNet (He et al. 2016), DenseNet (Huang et al. 2017), etc. Some of these architectures have become the state-of-the-art in computer vision applications such as face recognition, emotion detection, object extraction, scene classification, and also in conservation applications (e.g., species identification in camera trap data, land-use monitoring in aerial surveys). Given the image-like nature of time-frequency representations of acoustic signals (e.g., spectrogram), many of the successes of CNNs in computer vision have been replicated in the field of animal bioacoustics. In contrast to CNNs, RNNs are better suited for processing sequence inputs. RNNs contain internal states (memory) that allow them to "learn" temporal patterns. However, their utility is limited by the "vanishing gradient problem," wherein the gradients (from the gradient descent algorithm) of the network's output with respect to the weights in the early layers become extremely small. The problem is overcome in modern flavors of RNNs such as long short-term memory (LSTM; Hochreiter and Schmidhuber 1997) networks and gated recurrent unit (GRU; Cho et al. 2014) networks.

These types of ML solutions are heavily data-driven and often require large quantities of training samples. Typically, the training samples are time-frequency representations (e.g., spectrogram or mel-spectrogram) of short clips of recordings (e.g., Stowell et al. 2016; Shiu et al. 2020). Robustness of the resulting models are improved by ensuring that the inputs adequately cover possible variations of the target signals and of the ambient background conditions. Data scientists employ a variety of data augmentation techniques to overcome data shortage. Some examples include introducing synthetic variations such as infusion of Gaussian noise, shifting in time (horizontal shift) and frequency content (vertical shift) (Jaitly and Hinton 2013; Ko et al. 2015; Park et al. 2019). The training process, which involves successively lowering a loss function iteratively using the backpropagation algorithm, is usually computationally intensive and is often sped up with the use of GPUs.

DNNs have been used in the automatic recognition vocalizations of insects (e.g., Madhusudhana et al. 2019), fish (e.g., Malfante et al. 2018), birds (e.g., Stowell et al. 2016; Goëau et al. 2016), bats (e.g., Mac Aodha et al. 2018), marsupials (e.g., Himawan et al. 2018), primates (e.g., Zhang et al. 2018), and marine mammals (e.g., Bergler et al. 2019). CNNs have been used in the recognition of social calls, song calls, and whistles (e.g., Jiang et al. 2019; Thomas et al. 2019). While typical 2-dimensional CNNs have

been successfully used in the detection of echolocation clicks (e.g., Bermant et al. 2019), 1-dimensional CNNs (with waveforms as inputs) have been attempted as well (e.g., Luo et al. 2019). CNNs and LSTM networks have been compared in an application for classifying grouper species (Ibrahim et al. 2018) where the authors observed similar performances between the two models. Shiu et al. (2020) attempted combining a CNN with a GRU network for detecting North Atlantic right whale (*Eubalaena glacialis*) calls. Madhusudhana et al. (2021) incorporated long-term temporal context by combining independently trained CNNs and LSTM networks and achieved notable improvements in recognition performance. An attractive approach for developing recognition models is the use of transfer learning technique (Torrey and Shavlik 2010), where components of an already trained model are reused. Typically, weights of the early layers of a pre-trained network are frozen (no longer trainable) and the model is adapted to the target domain by training only the leaf nodes with data from the target domain. Zhong et al. (2020) used transfer learning to produce a CNN model for classifying the calls of a few species of frogs and birds.

#### 8.4.3.2 Random Forest Analysis

A random forest is a collection of many (hundreds or thousands) individual classification trees, which are grown without pruning. Each tree is different from every other tree in the forest because at each node, the variable to be used as a splitter is chosen from a random subset of the variables (Breiman 2001). Each tree in the forest produces a predicted category for the sound to be classified as, and the sound is ultimately classified as the category that was predicted by the majority of trees. Random forests are often more accurate than single classification trees because they are robust to over-fitting and stable to small perturbations in the data, correlations between predictor variables, and noisy predictor variables. Random forests perform well on polymorphic categories such as the variety of flight calls produced by many bird species (e.g., Liaw and Wiener 2002; Cutler et al. 2007; Armitage and Ober 2010; Ross and Allen 2014).

One of the advantages of a random forest analysis is that it provides information on the degree to which each one of the input variables contributes to the final species classification. This information is given by the Gini index and is known as the Gini variable importance. The Gini index is calculated based on the "purity" of each node in each of the classification trees, where purity is a measure of the number of whistles from different species in a given node (Breiman et al. 1984). Smaller Gini indices represent higher purity. When a random forest analysis is run, the algorithm assigns splitting variables so that the Gini index is minimized at each node (Oh et al. 2003). When a forest has been grown, the Gini importance value is calculated for each variable by summing the decreases in Gini index from one node to the next each time the variable is used. Variables are ranked according to their Gini importance values—those with the highest values contribute the most to the random forest model predictions. Random forests also produce a proximity measure, which is the fraction of trees in which particular observations end up in the same terminal nodes. This measure provides information about the similarity of individual observations because similar observations should end up in the same terminal nodes more often than dissimilar observations (Liaw and Wiener 2002).

Armitage and Ober (2010) compared the classification performance of random forests, support vector machines (SVMs), artificial neural networks, and DFA for bat echolocation signals and found that, with the exception of DFA, which had the lowest classification accuracy, all classifiers performed similarly. Keen et al. (2014) compared the performance of four classification algorithms using spectrographic measurements (spectrographic cross-correlation, dynamic time-warping, Euclidean distance, and random forest) for flight calls from four warbler species. In this study, random forests produced the most accurate results, correctly classifying 68% of calls.

Oswald et al. (2013) compared classifiers generated using DFA versus random forest classifiers for whistles produced by eight delphinid species recorded in the tropical Pacific Ocean and found that random forests resulted in the highest overall correct classification score. Rankin et al. (2016) trained a random forest classifier for five delphinid species in the California Current ecosystem. This classifier used information from whistles, clicks, and burst-pulse sounds and correctly classified 84% of acoustic encounters. Both Oswald et al. (2013) and Rankin et al. (2016) used spectrographic measurements as input variables for their classifiers.

#### 8.4.3.3 Gaussian Mixture Models

Gaussian Mixture Models (GMMs) are used commonly to model arbitrary distributions as linear combinations of parametric variables. They are appropriate for species identification when there are no expectations, such as the sequence of sounds (Roch et al. 2007). To create a GMM, a set of $n$ normal distributions with separate means and diagonal covariance matrices are scaled by weight-factors $c_i$ ($1 < i < n$). The sum over all $c_i$ must be 1 to ensure that the GMM represents a probability distribution (Huang et al. 2001; Roch et al. 2007, 2008). The number of mixtures in the GMM is chosen empirically and its parameters are estimated using an iterative algorithm, such as the Expectation Maximization algorithm (Moon 1996). Once a GMM has been trained, likelihood is computed for each sound type and a log-likelihood-ratio test is used to decide the species (Roch et al. 2008).

Gingras and Fitch (2013) used GMMs to classify male advertisement songs of four genera of anurans (*Bufo*, *Hyla*, *Leptodactylus, Rana*) based on spectral features and mel-frequency cepstral coefficients. The GMM based on spectral features resulted in 60% true positives and 13% false positives, and the GMM based on mel-frequency cepstral coefficients resulted in 41% true positives and 20% false positives. Somervuo et al. (2006) correctly classified 55–71% of song fragments from 14 different species of birds based on mel-frequency cepstral coefficients. The correct classification score depended on the number of cepstral coefficients and the number of Gaussian mixtures in the model. Lee et al. (2013) used GMMs to classify song segments of 28 species of birds based on image-shape features instead of traditional spectrographic features. This approach resulted in 86% or 95% classification accuracy for 3- or 5-s birdsong segments, respectively.

Roch et al. (2008) classified clicks produced by Blainville's beaked whales, pilot whales, and Risso's dolphins using a GMM. Correct classification scores for these three species were 96.7%, 83.2%, and 99.9%, respectively. Brown and Smaragdis (2008, 2009) used GMMs to classify sounds of killer whales, resulting in up to 92% agreement with 75 perceptually created categories of sound types, depending on the number of cepstral coefficients and Gaussians in the estimate of the probability density function. GMMs were used to classify the A and B type sounds produced by blue whales in the Northeast Pacific (McLaughlin et al. 2008), and six marine mammal species (Mouy et al. 2008) recorded in the Chukchi Sea: bowhead whales, humpback whales, gray whales, beluga whales, killer whales, and walruses. Both studies reported that their classifiers worked very well, but correct classification scores were not provided.

#### 8.4.3.4 Support Vector Machines

Support vector machines (SVMs) are a rich family of learning algorithms based on Vapnik's (1998) statistical learning theory. An SVM works by mapping features measured from sounds into a high-dimensional feature space. The SVM then finds the optimal hyperplane (function) that maximizes the separation among classes with the lowest number of parameters and the lowest risk of error. This approach attempts to meet the goal of minimizing both the training error and the complexity of the classifier (Mazhar et al. 2007). The best hyperplane is one that maximizes the distance between the hyperplane and the nearest data points belonging to different classes. The support vectors are the data points that determine the position of the hyperplane, and the distance between the hyperplane and the support vectors is called the margin (Fig. 8.21). The

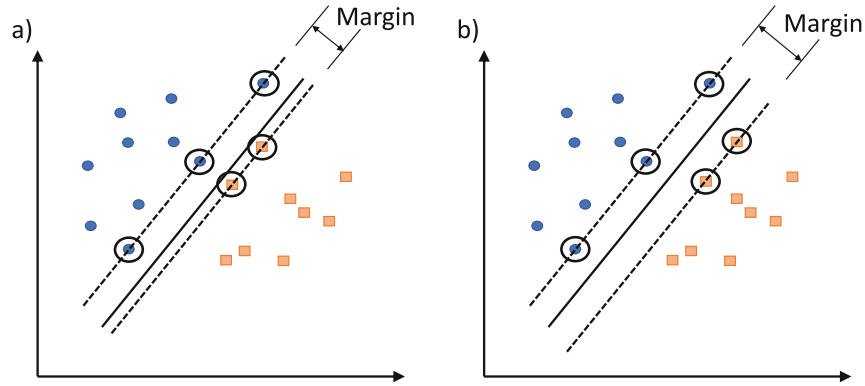


**Fig. 8.21** Examples of support vector machine hyperplanes. (**a**) The margin of the hyperplane is not optimal, (**b**) a hyperplane with a maximized margin. The support vectors are circled

optimal classifier maximizes the margin on both sides of the hyperplane. Because the hyperplane can be defined by only a few of the training samples, SVMs tend to be generalized and robust (Cortes and Vapnik 1995; Duda et al. 2001). When classes cannot be separated linearly, SVMs can map features onto a higher dimensional space where the samples become linearly separable (see Fig. 8.26 in Zeppelzauer et al. 2015).

SVMs originally were designed for binary classification, but a number of methods have been developed for applying them to multi-class problems. The three most common methods are: (1) form $k$ binary "one-against-the-rest" classifiers, where $k$ is the number of classes and the class whose decision-function is maximized is chosen (Vapnik 1998), (2) form all $k(k-1)/2$ pair-wise binary classifiers, and choose the class whose pair-wise decision-functions are maximized (Li et al. 2002), and (3) reformulate the objective function of SVM for the multi-class case so decision boundaries for all classes are optimized jointly (Guemeur et al. 2000).

Gingras and Fitch (2013) used four different algorithms (SVM, k-nearest neighbor, multivariate Gaussian distribution classifier, and GMM) to classify advertisement calls from four genera of anurans and obtained comparable accuracy levels from all three models. Fagerlund (2007) used SVMs to classify bird sounds produced by several species using decision trees with binary SVM classifiers at each node. The two datasets used by Fagerlund (2007) contained six and eight bird species and correct classification scores were 78–88% and 96–98% for the two datasets, respectively, depending on which variables were used in the classifiers.

Zeppelzauer et al. (2015) and Stoeger et al. (2012) both used SVM to identify African elephant rumbles. Zeppelzauer et al. (2015) used cepstral feature vectors and an SVM to distinguish African elephant rumbles from background noise. This SVM resulted in an 88% correct detection rate and a 14% false alarm rate. In addition to SVM, Stoeger et al. (2012) also used linear discriminant analysis (LDA) and nearest neighbor classification algorithms to categorize two types of rumbles produced by five captive African elephants based on spectral representations of the sounds. They obtained a classification accuracy of greater than 97% for all three classification methods.

Jarvis et al. (2006) developed a new type of multi-class SVM, called the class-specific SVM (CS-SVM). In this method, $k$ binary SVMs are created, where each SVM discriminates between one of the $k$ classes of interest and a common reference-class. The class whose decision-function is maximized with respect to the reference-class is selected. If all decision-functions are negative, the reference-class is selected. The advantage of this method is that noise in recordings is treated as the reference-

class. Jarvis et al. (2006) used their CS-SVM to discriminate clicks produced by Blainville's beaked whales from ambient noise and obtained a correct classification score of 98.5%. They also created a multi-class CS-SVM that classified clicks produced by Blainville's beaked whales, spotted dolphins (*Stenella attenuata*), and human-made sonar pings. This CS-SVM resulted in 98% correct classification for Blainville's beaked whale clicks, 88% correct classification for spotted dolphin clicks, and 95% correct classification for sonar pings. It is important to note that the training data were included in their test data, which likely resulted in inflated correct classification scores.

#### 8.4.3.5 Dynamic Time-Warping

Dynamic time-warping (DTW) is a class of algorithms originally developed for automated human speech recognition (Myers et al. 1980). DTW is used to quantitatively compare time-frequency contours of different durations using variable extension and compression of the time axis (Deecke and Janik 2006; Roch et al. 2007). There are different DTW techniques (e.g., Itakura 1975; Sakoe and Chiba 1978; Kruskal and Sankoff 1983), but all are based on comparing a reference sound to a test sound. The test sound is stretched and compressed along its contour to minimize the difference between the shapes of the two contours. Restrictions can be placed on the amount of time-warping that takes place. For example, Buck and Tyack (1993) did not time-warp contours that differed by a factor of more than 2 in duration and assigned those contours a similarity score of zero. Deecke and Janik (2006) stated that contours could only be stretched or compressed up to a factor of 3 to fit the reference contour. In a DTW analysis, all individual contours are compared to all other contours and a similarity matrix is constructed. Sounds are clustered into categories based on the similarity matrix using methods such as k-nearest neighbor cluster analysis or ANNs (Deecke and Janik 2006; Brown and Miller 2007).

DTW has been used to classify bird sounds. Anderson et al. (1996) applied DTW to recognize individual song syllables for two species of songbirds: indigo buntings (*Passerina cyanea)* and zebra finches (*Taeniopygia guttata*). Their analysis resulted in 97% correct classification of stereotyped syllables and 84% correct classification of syllables in plastic song. It is important to note, however, that these results were obtained for song recorded from a single individual of each species in a controlled setting. Somervuo et al. (2006) performed DTW to classify bird song syllables produced by 14 different species. They compared two different methods for computing distance between syllables: (1) simple Euclidean distances between frequency-amplitude vectors, and (2) absolute distance between frequencies weighted by the sum of their amplitudes. Classification accuracy was low, at about 40–50%, depending on the species and the distance method used. They obtained higher classification success using classification methods such as hidden Markov models (HMM) and GMM based on song fragments, rather than on single syllables.

Buck and Tyack (1993) performed DTW to classify three signature whistles from each of five wild bottlenose dolphins recorded in Sarasota, Florida, USA, with 100% accuracy. Deecke and Janik (2006) used DTW to classify signature whistles produced by captive bottlenose dolphins. The DTW algorithm outperformed human analysts and other statistical methods tested by Janik (1999). DTW also was applied to classify stereotypical pulsed sounds produced by killer whales, both in captivity (Brown et al. 2006) and at sea (Deecke and Janik 2006; Brown and Miller 2007). In all of these studies, sounds were classified into categories that were identified perceptually by humans with very high correct classification scores.

Oswald et al. (2021) used dynamic time-warping and neural network analysis to group whistle contours produced by short- and long-beaked common dolphins (*Delphinus delphis* and *D. bairdii*) into categories. Many of the resulting categories were shared between the two species, but each species also produced a number of species-specific categories. Random forest analysis showed that whistles in species-specific categories could be classified to species with significantly higher accuracy than whistles

in shared categories. This suggests that not every whistle carries species information, and that specific whistle types play an important role in dolphin species identification.

#### 8.4.3.6 Hidden Markov Models

Hidden Markov mode (HMM) theory was developed in the late 1960s by Baum and Eagon (1967) and now is used commonly for human speech recognition (Rabiner et al. 1983, 1996; Levinson 1985; Rabiner 1989). To create an HMM, a vector of features is extracted from a signal at discrete time steps. The temporal evolution of these features from one state to the next is modeled by creating a transition matrix $M$, where $M_{ij}$ is the probability of transition from state $i$ to state $j$, and an emission matrix $E$, where $E_{\text{is}}$ is the probability of observing signal $s$ in state $i$ (Rickwood and Taylor 2008). A different HMM is created for each species in the dataset and a sound is classified by determining which of the HMMs has the highest likelihood of producing that particular set of signal states. Training HMMs requires significant amounts of computing, and proper estimation of the transition and output probabilities is of crucial importance (Makhoul and Schwarz 1995). Excellent tutorials on HMMs can be found in Rabiner and Juang (1986) and Rabiner (1989).

A significant advantage inherent to HMMs is their ability to model time and spectral variability simultaneously (Makhoul and Schwarz 1995). They are able to model time series that have subtle temporal structure and are efficient for modeling signals with varying durations by performing nonlinear, temporal alignment during both the training and classification processes (Clemins et al. 2005; Roch et al. 2007; Trifa et al. 2008). Using HMMs, complex models can be built to deal with complicated biological signals (Rickwood and Taylor 2008), but care must be taken when choosing training samples to obtain a high generalization ability. The performance of an HMM is influenced by the size of the training set, the feature extraction method, and the number of states in the model (Trifa et al. 2008). Recognition performance is also affected by noise (Trifa et al. 2008).

In addition to being successfully implemented in human speech recognition, HMMs have been used to classify the sounds produced by birds (Kogan and Margoliash 1998; Trawicki et al. 2005, Trifa et al. 2008, Adi et al. 2010), red deer (*Cervus elaphus*; Reby et al. 2006), African elephants (Clemins et al. 2005), common dolphins (Sturtivant and Datta 1997; Datta and Sturtivant 2002), killer whales (Brown and Smaragdis 2008, 2009); beluga whales (Clemins and Johnson 2005; Leblanc et al. 2008), bowhead whales (Mellinger and Clark 2000), and humpback whales (Suzuki et al. 2006). HMMs perform as well as, or better than, both GMMs and DTW (Weisburn et al. 1993; Kogan and Margoliash 1998) and are becoming more common in animal classification studies.

Adi et al. (2010) also used HMMs to examine individually distinct acoustic features in songs produced by ortolan buntings (*Emberiza hortulana*). They represented each song syllable using a 15-state HMM (Fig. 8.22). These HMMs then were connected to represent song types. The 14 most common song types were included in the analysis and correct classification ranged from 50% to 99%, depending on the song type. Overall, 90% of songs were correctly classified. Adi et al. (2010) used these results to illustrate the feasibility of using acoustic data to assess population sizes for these birds.

Reby et al. (2006) used HMMs to examine whether common roars uttered by red deer during the rutting season can be used for individual recognition. They recorded roar bouts from seven captive red deer and used HMMs to model roar bouts as successions of silences and roars. Each roar in the analysis was modeled as a succession of states of frequency components measured from the roars. Overall, the HMM correctly identified 85% of roar bouts to the individual deer, showing that roars were individually specific. Reby et al. (2006) also used HMMs to examine stability in this individuality over the rutting season. They did this by training an HMM using roar bouts recorded at the beginning of the rutting season and testing the model using roar bouts recorded later in the rutting season. Overall, 58% of roar bouts were classified correctly, suggesting that individual identification cues in roar bouts varied over time.

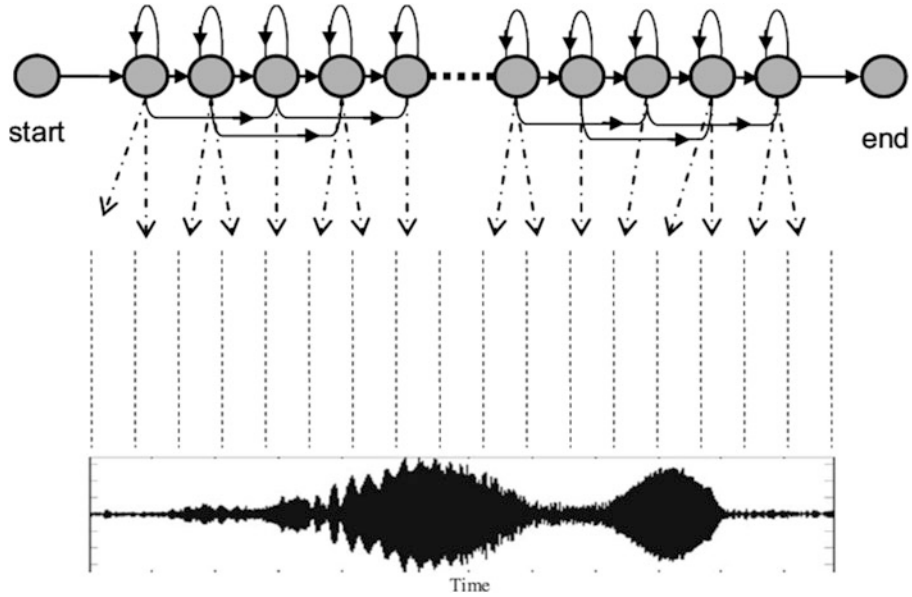


**Fig. 8.22** Example of a 15-state hidden Markov model representation of the waveform of a song syllable produced by an ortolan bunting to capture the temporal pattern of the syllable (Adi et al. 2010). © Acoustical Society of America, 2010. All rights reserved

## 8.5 Challenges in Classifying Animal Sounds

Placing sounds into categories is not always straightforward. Sounds produced by a particular species often contain a great deal of variability caused by different factors (e.g., location, date, age, sex, and individuality), which can make it difficult to define categories. In addition, sound categories are not always sharply demarcated, but instead grade or gradually transition from one form to another. It is important to be aware of the challenges in a particular dataset. Below are some types of variation that can be encountered in the classification of animal sounds.

### 8.5.1 Recording Artifacts

Bioacousticians need to be aware that recorded animal sounds are affected by the frequency and sensitivity specifications of the recording system used. An inappropriate recording system can result in distorted or partial sounds, which complicates their classification. For example, sounds can be misrepresented in recordings if the frequency response of the recording system is not linear, if the sampling frequency is too low, if sounds exist below or above the functional frequency range of the recording system, or if aliasing occurs (see Chap. 4). Ideally, recording systems should be carefully assembled and calibrated for the specific application. If the effects of the recording system could always be removed completely from recordings, sound classification would be more consistent and comparable. However, sounds published in the literature are sometimes received sounds that were affected by the recorder and/or the sound propagation environment.

One of the most common problems in underwater acoustic recordings is mooring noise. If hydrophones are held over the side of a boat, the recordings will contain sound from waves splashing against the boat or the hydrophone cable rubbing against the boat. Recorders built into mooring lines can record cable strum or clanking chains. If multiple oceanographic sensors are moored together, sounds from other

instruments (e.g., wipers on a turbidity sensor) may be recorded. Recorders resting on soft seafloor in coastal water may record the sound of sand swishing over the mooring. In addition, hydrostatic pressure fluctuations from the recorder bouncing in the water column or vortices at the hydrophone if deployed in strong currents will cause flow noise. All of these artifacts can last from seconds to minutes and appear in spectrograms as power from a few hertz to high kilohertz. Minimization of mooring noise and identification of recording artifacts is an art (also see Chaps. 2 and 3).

Similarly, artifacts can be recorded during airborne recordings. Wind is a primary artifact; however, moving vegetation and precipitation can also add noise to a recording. Any disturbance to the microphone can generate unwanted tapping or static on a recording. Recording systems in terrestrial environments need to be secured to minimize such noises.

### 8.5.2 Sound Propagation Effects

Environmental features of air or water can change the way sound propagates and thus the acoustic characteristics of a recorded sound. Bioacousticians need to understand environmental effects on the features of received sound to avoid classification of a signal variant as a new type, rather than as a particular sound type affected by propagation conditions. The sound propagation environment can affect both the spectral and temporal features of sound as it propagates from the animal to the recorder (see Chaps. 5 and 6). For example, energy at high frequencies is lost (attenuates) very quickly due to scattering and absorption, and therefore high-frequency harmonics do not propagate over long ranges. Acoustic energy at low frequencies (i.e., long wavelengths) does not travel well in narrow waveguides (e.g., shallow water). Because different frequencies within a sound can attenuate at different rates, the same sound can appear differently on a spectrogram, depending on the distance at which it was recorded.

Differential attenuation of frequencies in air is shown in Fig. 8.23. Signals produced by a big brown bat (*Eptesicus fuscus*) flying toward a microphone contain more ultrasonic components than signals recorded from a bat flying away from the microphone. The signal with the longest frequency modulation (from 100 to 50 kHz) is received when the bat is closest to the microphone. Variations in this spectrogram show how one sound type could be categorized differently simply because of distance between the animal and recorder, orientation to the microphone, and the gain setting.

Other sound propagation effects include reverberation (which leads to the temporal spreading of brief, pulsed sounds) and frequency dispersion. Frequency dispersion is a result of energy at different frequencies traveling at different speeds. This leads to sounds being spread out in time and, specifically in some underwater environments, can cause pulsed sounds to become frequency-modulated sounds (either up- or downsweeps; Fig. 8.24).

Finally, ambient noise (i.e., geophysical noise, anthropogenic noise, and non-target biological noise) superimposes with animal sounds, and at some distances and frequencies, parts of the animal sound spectrum will begin to drop below the levels of ambient noise. As a result, the same animal sound in a different environment and at a different distance from the animal can look quite different on a spectrogram and cause it to be misclassified as two different sound types.

### 8.5.3 Angular Aspects of Sound Emission

The orientation of an animal relative to the receiver (microphone or hydrophone) can change the acoustic features of the recorded sound. This complicates classification, and off-axis variations of a sound need to be known so they can be categorized as just a variant of a particular sound type, rather than as a new sound type. Not all sounds emitted by animals are omnidirectional (i.e., propagate equally in all angles relative to the animal). Au et al. (2012) studied the directionality of bottlenose dolphin echolocation clicks by measuring the horizontal and vertical emission beam patterns of these sounds. The angle at which an echolocation click was

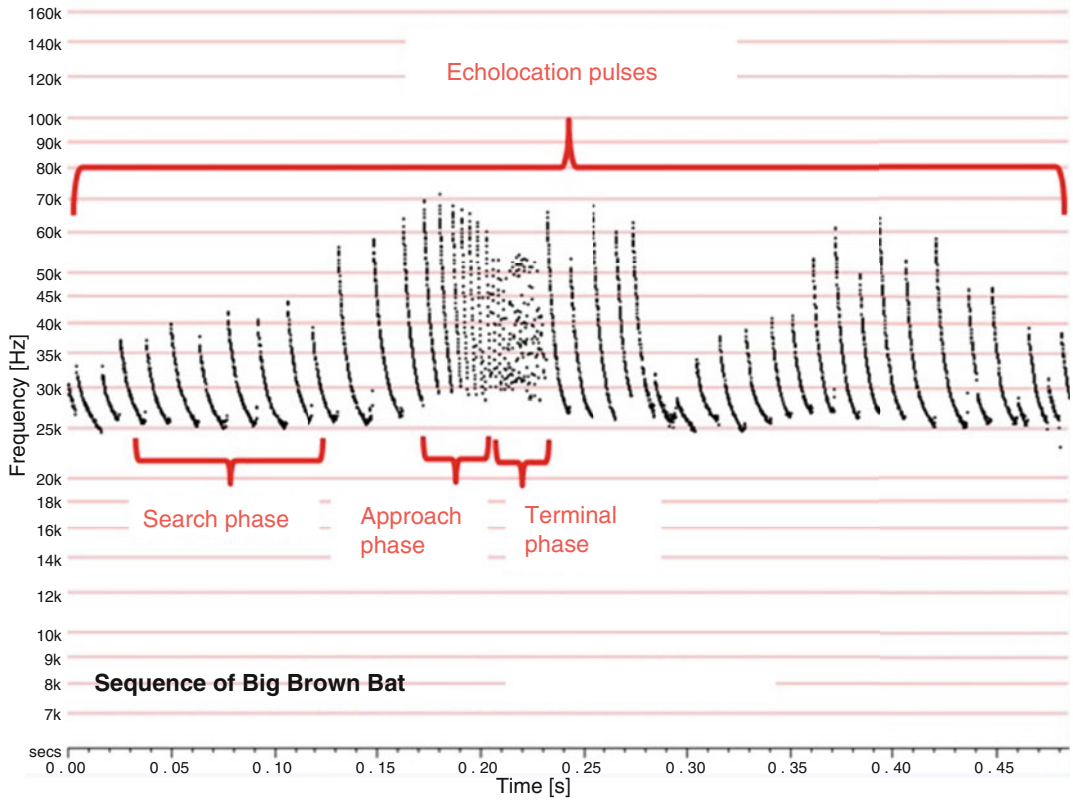


**Fig. 8.23** Spectrogram of big brown bat (*Eptesicus fuscus*) circling a recording device while searching and pursuing aerial prey. As the bat approaches the microphone, more of the ultrasonic signal is received (calls reach up to 70 kHz). As the bat moves away, the signal is attenuated. Time between calls shortens notably as the bat pursues an insect prey for capture. Notice that the bat emits "search" calls at 25–40 kHz, approach calls at 30–70 kHz when it is in pursuit or trying to navigate flight through complex space, and finally terminal calls at 30–55 kHz

recorded relative to the transducer (or echolocating animal) not only affected its received level, but also the waveform and frequency spectrum (Fig. 8.25). Sperm whale (*Physeter macrocephalus*) echolocation clicks, when recorded off-axis (i.e., away from the center of its emission beam), consisted of multiple complex pulses that were likely due to internal reflections within the sperm whale's head (Møhl et al. 2003; also see Chap. 12).

### 8.5.4 Geographic Variation

Geographic variation, or differences in the sounds produced by populations of the same species living in different regions, has been documented for many terrestrial and aquatic animals, including Hawaiian crickets (Mendelson and Shaw 2003), Túngara frogs (*Engystomops pustulosus*, Pröhl et al. 2006), bats (Law et al. 2002; Aspetsberger et al. 2003; Russo et al. 2007; Yoshino et al. 2008), pikas (Borisova et al. 2008), sciurid rodents (Gannon and Lawlor 1989; Slobodchikoff et al. 1998; Yamamoto et al. 2001; Eiler and Banack 2004), singing mice (*Scotinomys* spp., Campbell et al. 2010), primates (Mitani et al. 1992; Delgado 2007; Wich et al. 2008), cetaceans (Helweg et al. 1998; McDonald et al. 2006; Delarue et al. 2009; Papale et al. 2013, 2014), and elephant seals (*Mirounga* spp., Le Boeuf and Peterson

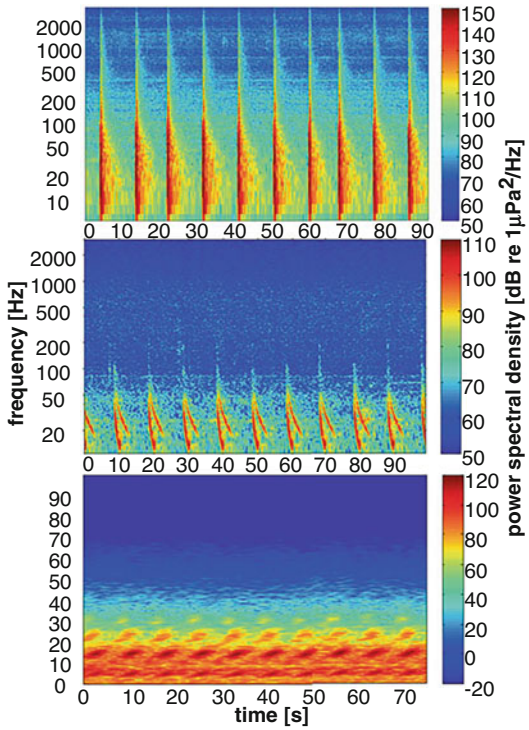


**Fig. 8.24** Spectrograms of marine seismic airgun signals recorded at three different ranges: 1.5 km (top), 80 km over soft seabed (middle), and 40 km over a hard seabed (bottom). The top and bottom spectrograms are of the same seismic survey. Pulses were brief and broadband near the source, but became frequency-modulated and narrowband some distance away due to dispersion (Erbe et al. 2016). © Erbe et al.; https://ars.els-cdn.com/content/image/1-s2.0-S0025326X15302125-gr9_lrg.jpg. Licensed under CC BY 4.0; https://creativecommons.org/licenses/by/4.0/

1969). When developing classifiers, it is important to understand the degree of geographic variation in a sound repertoire and the range over which this occurs. If geographic variation exists, then a classifier trained using data collected in one location may not work well when applied to data collected in another location.

One of the underlying causes of geographic variation may be reproductive isolation of a population. Keighley et al. (2017) used DFA with stepwise variable selection to determine geographic variation in sounds from six major populations of palm cockatoos (*Probosciger aterrimus*) in Australia. Palm cockatoos from the east coast (Iron Range National Park) had unique contact sounds and produced fewer sound types than at other locations. The authors speculated that this large difference was due to long-term isolation at this site and noted that documentation of geographic variation in sounds provided important conservation information for determining connectivity of these six populations.

Thomas and Golladay (1995) employed PCA to classify nine underwater vocalization types produced by leopard seals (*Hydrurga leptonyx*) at three study sites near Palmer Peninsula, Antarctica. The PCA successfully separated vocalizations from the three study areas and provided information about what features of the sounds were driving the differences among locations. For example, the first principal component was influenced by maximum, minimum, start, and end frequencies, the second principal component was influenced by the presence or absence of overtones, and the third principal component was predominantly related to time relationships, such as duration and time between successive sounds. Note that some sound types were absent at some locations.

### 8.5.5 Graded Sounds

Some animals produce sound types that grade or gradually transition from one type to another. Researchers should not neglect the potential existence of vocal intermediates in classification. For example, Schassburger (1993) described sounds produced by timber wolves (*Canis lupus*) as barks, growl-moans, growls, howls moans, snarls, whimpers, whine-moans, whines, woofs, and yelps. Wolves combine these 11 principal sounds to create mixed-sounds that often grade from one type into another.

Clicks trains, burst-pulse sounds, and whistles produced by delphinids are typically considered as three distinct categories of sound. Click trains and burst-pulse sounds are composed of short, exponentially damped sine waves separated by periods of silence, while whistles are generally thought of as continuous tonal sounds, often

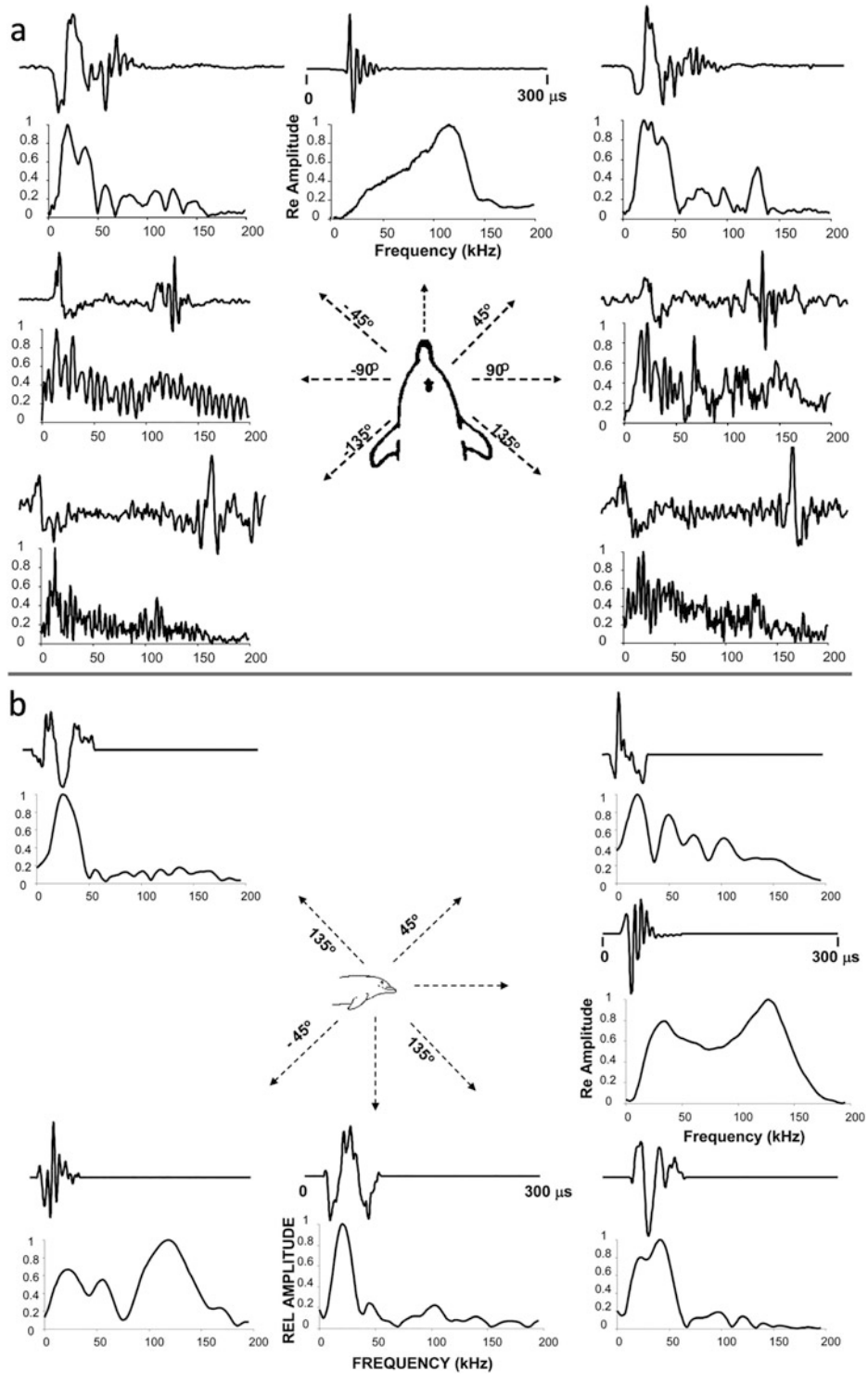


**Fig. 8.25** Waveforms and spectra of a bottlenose dolphin echolocation click in the horizontal (**a**) and vertical (**b**) planes (Au et al. 2012). © Acoustical Society of America, 2012. All rights reserved

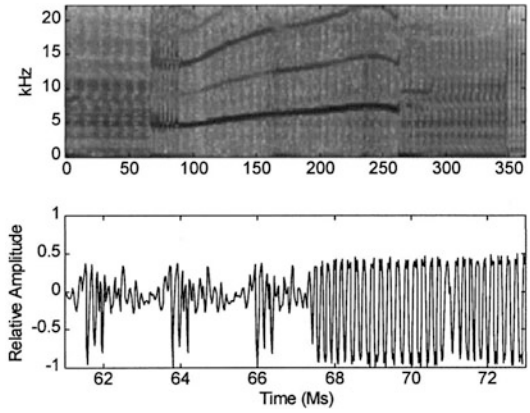


**Fig. 8.26** Spectrogram and waveform of a false killer whale vocalization. The vocalization appears to be a whistle in the spectrogram, but the waveform reveals discrete pulses between 61 and 67 ms (Murray et al. 1998). © Acoustical Society of America, 1998. All rights reserved

sweeping in frequency. While these sounds appear quite different from one another on spectrograms, closer inspection of their waveforms reveals that some sounds that look like whistles on a spectrogram actually contain a high degree of amplitude modulation. In other words, some sounds that are considered to be whistles are made up of pulses with inter-pulse intervals that are too short to hear or be resolved by the analysis window of the spectrogram (Fig. 8.26). As an example of this, Murray et al. (1998) used self-organizing neural networks to analyze the vocal repertoires of two captive false killer whales (*Pseudorca crassidens*) based on measurements taken from waveforms. They found that rather than organizing sounds into distinct categories, the vocal repertoire was more accurately represented by a graded continuum, with exponentially damped sinusoidal pulses on one end and continuous sinusoidal signals at the other. Beluga whales also have been shown to have a graded vocal repertoire (Karlsen et al. 2002; Garland et al. 2015). Whistles with a high degree of amplitude modulation have been recorded from Atlantic spotted and spinner (*Stenella longirostris*) dolphins (Lammers et al. 2003), suggesting that this graded continuum model is applicable to these species as well.

### 8.5.6 Repertoire Changes Over Time

Some animal sound repertoires change over time, which complicates their classification. For example, humpback whale song slowly changes over the course of a breeding season as new units are introduced and old ones discarded (Noad et al. 2000). Song also changes from one season to the next, and in one instance, eastern Australian humpback whales changed to the song of the western Australian population within 1 year (Noad et al. 2000).

Antarctic blue whales can be heard off southwestern Australia from February to October every year. The upper frequency of their Z-call decreases over the season by about 0.4–0.5 Hz. At the beginning of the next season, the Z-call jumps in frequency to about the mean of the Z frequency of the previous season, and then decreases again, leading to an average decrease in the frequency of the upper part of the Z-call by $0.135 \pm 0.003$ Hz/year (Fig. 8.27; Gavrilov et al. 2012). A similar decrease (albeit at different rates at different locations) has been observed for the "spot call," of which the animal source remains elusive (Fig. 8.27; Ward et al. 2017). The reasons for these shifts are unknown.

## 8.6 Summary

Animals, whether they are in air, on land, or under water, produce sound in support of their various life functions. Cicadas join in chorus to repel predatory birds (Simmons et al. 1971); male fishes chorus on spawning grounds to attract females (Amorim et al. 2015); frogs call to attract mates and to mark out their territory (Narins et al. 2006); birds, too, sing for territorial and reproductive reasons (Catchpole and Slater 2008); bats emit clicks for echolocation during hunting and navigating, as do dolphins (Madsen and Surlykke 2013). In order to study animals by listening to their sounds, sounds need to be classified to species, to behavior, etc. In the early days, this was done without measurements or with only the simplest measuring tools. Scientists listened to the

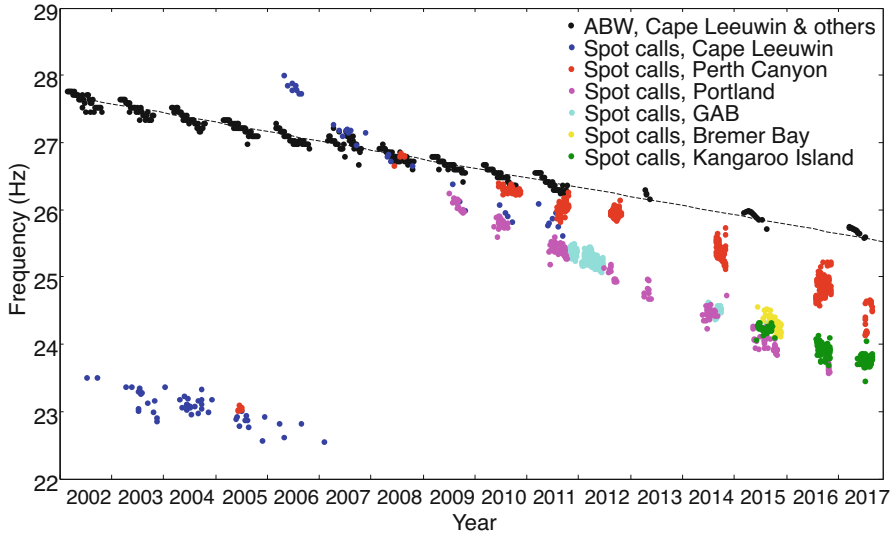


**Fig. 8.27** Weekly means of the upper part of the Antarctic blue whale Z-call over several years, as well as of the spot call, which remains to be identified to species. All locations are off Australia (GAB: Great Australian Bight). Data updated from Gavrilov et al. (2012) and Ward et al. (2017). Courtesy of Sasha Gavrilov

sounds in the field, often while visually observing animals. Scientists recorded sounds in the field and analyzed the recordings in the laboratory by listening, looking at oscillograms or spectrograms, and manually sorting sounds into types. Nowadays, with the affordability of autonomous recording equipment, bioacousticians collect vast amounts of data, which can no longer be analyzed without the aid of automated data processing, data reduction, and data analysis tools. Given simultaneous advances in computer hard- and software, datasets may be analyzed more efficiently, and with the added advantage of reducing opportunities for human subjective biases.

In this chapter, we presented software tools for automatically detecting animal sounds in acoustic recordings, and for classifying those sounds. The detectors we discussed compute a specific quantity of the sound (such as its instantaneous energy or entropy) and then apply a threshold above which the sound is deemed detected. The specific detectors were based on acoustic energy, Teager–Kaiser energy, entropy, matched filtering, and spectrogram cross-correlation. Setting the detection threshold critically affects how many signals are detected and how many are missed. We presented two ways of finding the best threshold and assessing detector performance: receiver operating characteristics and precision-recall curves.

Once signals have been detected, they can be classified. A common pre-processing step immediately prior to classification includes the measurement of sound features such as minimum and maximum frequency, duration, or cepstral features. The software tools we presented for classification included parametric clustering, principal component analysis, discriminant function analysis, classification trees, and machine learning algorithms. No single tool outperforms all others; rather, the best tool suited for the specific task needs to be employed. We discussed advantages and limitations of the various tools and provided numerous examples from the literature. Finally, challenges resulting from recording artifacts, the environment affecting sound features, and changes in sound features over time and space were explored.

It is important to remember that human perception of a sound likely is not the same as an animal's perception of the sound and yet

bioacousticians commonly describe or classify animal sounds in human terms. Classification of the acoustic repertoire of an animal into sound types provides a convenient framework for comparing and contrasting sounds, taking systematic measurements from portions of the repertoire, and performing statistical analyses. However, categories determined based on human perception may have little or no relevance to the animals and so human categorizations can be biologically meaningless. For example, humans have limited low-frequency and high-frequency hearing abilities compared to many other species, and so aural classification of sound types is sometimes based on only a portion of a sound audible to the human listener. Whether sound types determined by humans are meaningful classes to the animals is mostly unknown. While categorizing sounds based on function is an attractive approach for the behavioral zoologist, establishing the functions of these sounds is often challenging. In our review of classification methods, it was clear that methods developed for human speech could be applied to animal sounds. Some fascinating questions lie ahead for bioacousticians as they attempt to extend understanding of the perception experienced by other animals.

Even with the above caveats, detection and classification of animal sounds is useful for research and conservation. It allows populations to be monitored, their distribution and abundance to be determined, and impacts (e.g., from human presence or climate change) to be assessed. It can also be useful for conservation of a species (i.e., to create taxonomy, identify geographic variation in populations, examine ecological connectivity among populations, and detect changes in the biological uses sounds due to the advent and growth of anthropogenic noise). Classification of animal sounds is important for understanding behavioral ecology and social systems of animals and can be used to identify individuals, social groups, and populations. The ability to study these types of topics will ultimately lead to a deeper understanding of the evolutionary forces that shape animal bioacoustics.

With a goal to foster wider participation in research on bioacoustic pattern recognition, a number of global competitions are held regularly. The annual Detection and Classification of Acoustic Scenes and Event (DCASE) workshops and BirdCLEF challenges (part of Cross Language Evaluation Forum) attract hundreds of data scientists for developing machine learning solutions for recognizing bird sounds in soundscape recordings. The marine mammal community organizes the biennial Detection, Classification, Localization, and Density Estimation (DCLDE) workshops. These challenges put out large training datasets for researchers to develop detection and classification systems, assess the performance of submitted solutions with "held out" datasets, and reward the top-ranked submissions. The datasets from these challenges are often made available for use by the research community after the competitions, while some workshops make available the submitted solutions as well.

## 8.7 Additional Resources

- *PAMGuard* is an open-source software package for acoustic detection, classification, and localization of cetacean sounds: https://www.pamguard.org/
- *Ishmael* is a free software package for acoustic detection, classification, and localization of cetacean sounds: http://www.bioacoustics.us/ishmael.html
- *Koe* is a free, web-based software for annotation, measurement, and classification of bioacoustics signals: https://koe.io.ac.nz/# (Fukuzawa et al. 2020)
- *Praat* is free software originally designed for human speech analysis, but used by many bioacousticians: https://www.fon.hum.uva.nl/praat/
- *Characterization Of Recorded Underwater Sound (CHORUS)* is a *MATLAB* graphic user interface developed by Curtin University, Perth, WA, Australia, with built-in automatic detectors for pygmy blue and fin whales

(Gavrilov and Parsons 2014): https://cmst.curtin.edu.au/products/chorus-software/
- Detection, Classification, Localization, and Density Estimation of Marine Mammals using Passive Acoustics meeting websites:
  - Mount Hood, Oregon, USA, 2011: http://www.bioacoustics.us/dcl.html
  - St Andrews, Scotland, UK, 2013: https://soi.st-andrews.ac.uk/dclde2013/
  - San Diego, California, USA, 2015: http://www.cetus.ucsd.edu/dclde/index.html
  - Paris, France, 2018: http://sabiod.univ-tln.fr/DCLDE/
  - Hawaii, USA, 2022: http://www.soest.hawaii.edu/ore/dclde/
- Bird sound recognition challenges: http://dcase.community/ (DCASE), https://www.imageclef.org/BirdCLEF2020 (BirdCLEF)
- *BirdNET* is an Android app for birdsong recognition: https://birdnet.cornell.edu/
- *SongSleuth* is an Apple or Android app for birdsong recognition: https://www.songsleuth.com/#/
- All accessed 5 Aug 2022.

## References


Adi K, Johnson MT, Osiejuk TS (2010) Acoustic censusing using automatic vocalization classification and identity recognition. J Acoust Soc Am 127:874–883. https://doi.org/10.1121/1.3273887

Afifi AA, Clark V (1996) Computer-aided multivariate analysis, 3rd edn. Chapman and Hall/CRC, New York

Amorim MC, Vasconcelos RO, Fonseca PJ (2015) Fish sounds and mate choice. In: Ladich F (ed) Sound communication in fishes. Springer, Vienna, pp 1–33

Anderson SE, Dave AS, Margoliash D (1996) Template-based automatic recognition of birdsong syllables from continuous recordings. J Acoust Soc Am 100:1209–1219. https://doi.org/10.1121/1.415968

Armitage DW, Ober HK (2010) A comparison of supervised learning techniques in the classification of bat echolocation calls. Ecol Inform 5:465–473. https://doi.org/10.1016/j.ecoinf.2010.08.001

Aspetsberger F, Brandsen D, Jacobs DS (2003) Geographic variation in the morphology, echolocation and diet of the little free-tailed bat, *Chaerephon pumilus* (Molossidae). Afr Zool 38:245–254. https://doi.org/10.1080/15627020.2003.11407278

Au WWL, Nachtigall PE (1995) Artificial neural network modeling of dolphin echolocation. In: Kastelein RA, Thomas JA, Nachtigall PE (eds) Sensory systems of aquatic mammals. De Spil Publishers, Woerden, The Netherlands, pp 183–199

Au WWL, Branstetter B, Moore P, Finneran J (2012) The biosonar field around an Atlantic bottlenose dolphin (*Tursiops truncatus*). J Acoust Soc Am 131(1): 569–576. https://doi.org/10.1121/1.3662077

Baptista LF, Gaunt SSL (1997) Social interaction and vocal development in birds. In: Snowden CT, Hausberger M (eds) Social influences on vocal development. Cambridge Univ Press, Cambridge, pp 23–40

Baum LE, Eagon JA (1967) An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bull Am Math Soc 73:360–363

Baumgartner MF, Fratantoni DM (2008) Diel periodicity in both Sei whale vocalization rates and the vertical migration of their copepod prey observed from ocean gliders. Limnol Oceanogr 53:2197–2209. https://doi.org/10.4319/lo.2008.53.5_part_2.2197

Beeman K (1998) Digital signal analysis, editing and synthesis. In: Hopp SL, Owren MJ, Evans CS (eds) Animal acoustic communication: sound analysis and research methods. Springer, Berlin, pp 59–103

Belliustin NS, Kuznetsov SO, Nuidel IV, Yakhno VG (1991) Neural networks with close nonlocal coupling for analyzing composite image. Neurocomputing 3: 231–246. https://doi.org/10.1016/0925-2312(91)90005-V

Bergler C, Schröter H, Cheng RX, Barth V, Weber M, Nöth E, Hofer H, Maier A (2019) ORCA-SPOT: an automatic killer whale sound detection toolkit using deep learning. Sci Rep 9(1):1–7. https://doi.org/10.1038/s41598-019-47335-w

Bermant PC, Bronstein MM, Wood RJ, Gero S, Gruber DF (2019) Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. Sci Rep 9(1):1–10. https://doi.org/10.1038/s41598-019-48909-4

Borisova NG, Rudneva LV, Starkov AI (2008) Interpopulation variability of vocalizations in the Daurian pika (*Ochotona daurica*). Zool Zh 87:850–861

Bouffaut L, Dréo R, Labat V, Boudraa AO, Barruol G (2018) Passive stochastic matched filter for Antarctic blue whale call detection. J Acoust Soc Am 144(2): 955–965. https://doi.org/10.1121/1.5050520

Bradbury JW, Vehrencamp SL (2011) Principles of animal communication, 2nd edn. Sinauer Associates, New York

Brandes TS (2008) Feature-vector selection and use with Hidden Markov Models to identify frequency-modulated bioacoustic signals amidst noise. IEEE Trans Speech Lang Process 16:1173–1180. https://doi.org/10.1109/TASL.2008.925872

Breiman L (2001) Random forests. Mach Learn 45:5–32

Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth, Pacific Grove, CA

Briefer EF, Maigrot A-L, Roi T, Mandel R, Briefer Freymond S, Bachmann I, Hillmann E (2015) Segregation of information about emotional arousal and

valence in horse whinnies. Sci Rep 5(1):1–11. https://doi.org/10.1038/srep09989
Briskie JV, Martin PR, Martin TE (1999) Nest predation and the evolution of nestling begging calls. Proc R Soc Lond B 266:2153–2159. https://doi.org/10.1098/rspb.1999.0902
Brown JC, Miller PJO (2007) Automatic classification of killer whale vocalizations using dynamic time warping. J Acoust Soc Am 122:1201–1207. https://doi.org/10.1121/1.2747198
Brown JC, Smaragdis P (2008) Automatic classification of vocalizations with Gaussian mixture models and Hidden Markov Models. J Acoust Soc Am 123:3345. https://doi.org/10.1121/1.2933896
Brown JC, Smaragdis P (2009) Hidden Markov and Gaussian mixture models for automatic sound classification. J Acoust Soc Am 125:EL221–EL224. https://doi.org/10.1121/1.3124659
Brown JC, Hodgins-Davis A, Miller PJO (2006) Classification of vocalizations of killer whales using dynamic time warping. J Acoust Soc Am 119:EL34–EL40. https://doi.org/10.1121/1.2166949
Buck JR, Tyack PL (1993) A quantitative measure of similarity for *Tursiops truncatus* signature whistles. J Acoust Soc Am 94:2497–2506. https://doi.org/10.1121/1.407385
Camacho-Alpízar A, Fuchs EJ, Barrantes G (2018) Effect of barriers and distance on song, genetic, and morphological divergence in the highland endemic Timberline Wren (*Thryorchilus browni*, Troglodytidae). PLoS One 13(12):e0209508. https://doi.org/10.1371/journal.pone.0209508
Campbell P, Pasch B, Pino JL, Crino OL, Phillips M, Phelps SM (2010) Geographic variation in the songs of neotropical singing mice: testing the relative importance of drift and local adaptation. Evolution 64(7):1955–1972. https://doi.org/10.1111/j.1558-5646.2010.00962.x
Catchpole CK, Slater PJB (2008) Bird song: biological themes and variations, 2nd edn. Cambridge University Press, Cambridge
Cerchio S, Jacobsen JK, Norris TF (2001) Temporal and geographical variation in songs of humpback whales, *Megaptera novaeangliae*: synchronous change in Hawaiian and Mexican breeding assemblages. Anim Behav 62(2):313–329. https://doi.org/10.1006/anbe.2001.1747
Cho K, Van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: encoder-decoder approaches. arXiv:1409.1259
Clark CW (1980) A real-time direction-finding device for determining the bearing to the underwater sounds of southern right whales, *Eubalaena australis*. J Acoust Soc Am 68:508–511. https://doi.org/10.1121/1.384762
Clark CW (1982) The acoustic repertoire of the southern right whale, a quantitative analysis. Anim Behav 30(4):1060–1071. https://doi.org/10.1016/S0003-3472(82)80196-6
Clark LA, Pregibon D (1992) Statistical models. In: Chambers SJM, Hastie TJ (eds) Statistical models in S. Wadsworth and Brooks/Cole, Pacific Grove, CA
Clarke E, Reichard UH, Zuberbühler K (2006) The syntax and meaning of wild gibbon songs. PLoS One 1(1):E73. https://doi.org/10.1371/journal.pone.0000073
Clemins PJ, Johnson MT (2005) Unsupervised classification of beluga whale vocalizations. J Acoust Soc Am 117:2470. https://doi.org/10.1121/1.4809461
Clemins PJ, Johnson MT, Leong KM, Savage A (2005) Automatic classification and speaker identification of African elephant (*Loxodonta africana*) vocalizations. J Acoust Soc Am 117:956–963. https://doi.org/10.1121/1.1847850
Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker SW (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc Natl Acad Sci 102(21):7426–7431. https://doi.org/10.1073/pnas.0500334102
Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297
Courts R, Erbe C, Wellard R, Boisseau O, Jenner KC, Jenner M-N (2020) Australian long-finned pilot whales (*Globicephala melas*) emit stereotypical, variable, biphonic, multi-component, and sequenced vocalisations, similar to those recorded in the northern hemisphere. Sci Rep 10(1):20609. https://doi.org/10.1038/s41598-020-74111-y
Crance JL, Berchok CL, Wright DL, Brewer AM, Woodrich DF (2019) Song production by the North Pacific right whale, *Eubalaena japonica*. J Acoust Soc Am 145(6):3467–3479. https://doi.org/10.1121/1.5111338
Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88:2783–2792. https://doi.org/10.1890/07-0539.1
Dang T, Bulusu N, Hu W (2008) Lightweight acoustic classification for cane toad monitoring. In: 42nd Asilomar Conference on Signals, Systems and Computers. IEEE, New York, pp 1601–1605
Datta S, Sturtivant C (2002) Dolphin whistle classification for determining group identities. Sig Process 82(2):251–258. https://doi.org/10.1016/S0165-1684(01)00184-0
Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA
Davis SB, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Sig Process 28:357–366. https://doi.org/10.1109/TASSP.1980.1163420
Dawson MRW, Charrier I, Sturdy CB (2006) Using an Artificial Neural Network to classify black-capped chickadee (*Poecile atricapillus*) sound note types. J Acoust Soc Am 119(5):3161–3172. https://doi.org/10.1121/1.2189028

Deecke VB, Janik VM (2006) Automated categorization of bioacoustic signals: avoiding perceptual pitfalls. J Acoust Soc Am 119:645–653. https://doi.org/10.1121/1.2139067

Deecke VB, Ford JKB, Spong P (1999) Quantifying complex patterns of bioacoustic variation: use of a neural network to compare killer whale (*Orcinus orca*) dialects. J Acoust Soc Am 105:2499–2507. https://doi.org/10.1121/1.426853

Delarue J, Todd SK, Van Parijs SM, Di Iorio L (2009) Geographic variation in Northwest Atlantic fin whale (*Balaenoptera physalus*) song: implications for stock structure assessment. J Acoust Soc Am 125:1774–1782. https://doi.org/10.1121/1.3068454

Delgado RA (2007) Geographic variation in the long sounds of male orangutans (*Pongo* spp.). Ethology 113:487–498. https://doi.org/10.1111/j.1439-0310.2007.01345.x

Deregnaucourt S, Guyomarch JC, Richard V (2001) Classification of hybrid crows in quail using artificial neural networks. Behav Process 56:103–112. https://doi.org/10.1016/S0376-6357(01)00188-7

Duda R, Hart P, Stork D (2001) Pattern classification, 2nd edn. Wiley, Hoboken, NJ

Dunlop RA, Noad MJ, Cato DH, Stokes D (2007) The social vocalization repertoire of east Australian migrating humpback whales (*Megaptera novaeangliae*). J Acoust Soc Am 122(5):2893–2905. https://doi.org/10.1121/1.2783115

Dunlop RA, Cato DH, Noad MJ, Stokes DM (2013) Source levels of social sounds in migrating humpback whales (*Megaptera novaeangliae*). J Acoust Soc Am 134(1):706–714. https://doi.org/10.1121/1.4807828

Egan JP (1975) Signal detection theory and ROC analysis. Academic Press, New York

Eiler KC, Banack SA (2004) Variability in the alarm call of golden-mantled ground squirrels (*Spermophilus lateralis* and *S. saturatus*). J Mammal 85:43–50. https://doi.org/10.1644/1545-1542(2004)085<0043:VITACO>2.0.CO;2

Erbe C, King AR (2008) Automatic detection of marine mammals using information entropy. J Acoust Soc Am 124(5):2833–2840. https://doi.org/10.1121/1.2982368

Erbe C, Verma A, McCauley R, Gavrilov A, Parnum I (2015) The marine soundscape of the Perth Canyon. Prog Oceanogr 137:38–51. https://doi.org/10.1016/j.pocean.2015.05.015

Erbe C, Reichmuth C, Cunningham K, Lucke K, Dooling R (2016) Communication masking in marine mammals: a review and research strategy. Mar Pollut Bull 103:15–38. https://doi.org/10.1016/j.marpolbul.2015.12.007

Erbe C, Dunlop R, Jenner KCS, Jenner M-NM, McCauley RD, Parnum I, Parsons M, Rogers T, Salgado-Kent C (2017) Review of underwater and in-air sounds emitted by Australian and Antarctic marine mammals. Acoust Aust 45:179–241. https://doi.org/10.1007/s40857-017-0101-z

Esfahanian M, Erdol N, Gerstein E, Zhuang H (2017) Two-stage detection of north Atlantic right whale upcalls using local binary patterns and machine learning algorithms. Appl Acoust 120:158–166. https://doi.org/10.1016/j.apacoust.2017.01.025

Fagerlund S (2007) Bird species recognition using support vector machines. EURASIP J Appl Sig Proc 2007(1):1–8. https://doi.org/10.1155/2007/38637

Fenton MB, Jacobson SL (1973) An automatic ultrasonic sensing system for monitoring the activity of some bats. Can J Zool 51:291–299. https://doi.org/10.1139/z73-041

Fitch WT (2003) Mammalian vocal production: themes and variation. In: Proceedings of the 1st International Conference on Acoustic Communication by Animals, 27–30 July, pp 81–82

Forti LR, Costa WP, Martins LB, Nunes-de-Almeida CH, Toledo LF (2016) Advertisement call and genetic structure conservatism: good news for an endangered Neotropical frog. PeerJ 4:e2014. https://doi.org/10.7717/peerj.2014

Freitag LE, Tyack PL (1993) Passive acoustic localization of the Atlantic bottlenose dolphin using whistles and echolocation clicks. J Acoust Soc Am 93:2197–2205. https://doi.org/10.1121/1.406681

Fristrup KM, Watkins WA (1993) Marine animal sound classification. Woods Hole Oceanographic Institution Technical Report WHOI-94-13, p 29

Frommolt K-H, Bardeli R, Clausen M (eds) (2007) Computational bioacoustics for assessing biodiversity. Proceed Internat Expert meeting on IT-based detection of bioacoustical patterns, 7–10 December 2007 at the International Academy for Nature Conservation (INA) Isle of Vilm, Germany. BfN - Skripten Federal Agency for Nature Conservation, p 234

Fukushima K, Wake N (1990) Alphanumeric character recognition by neocognitron. In: Miller RE (ed) Advanced neural computers. Elsevier Science, Amsterdam, pp 263–270

Fukuzawa Y, Webb WH, Pawley MD, Roper MM, Marsland S, Brunton DH, Gilman A (2020) Koe: web-based software to classify acoustic units and analyse sequence structure in animal vocalizations. Methods Ecol Evol 11:431–441. https://doi.org/10.1111/2041-210X.13336

Gannier A, Fuchs S, Quebre P, Oswald JN (2010) Performance of a contour-based classification method for whistles of Mediterranean dolphins. Appl Acoust 7:1063–1069. https://doi.org/10.1016/j.apacoust.2010.05.019

Gannon WL, Lawlor TE (1989) Variation in the chip vocalization of three species of Townsend's chipmunks (genus *Eutamias*). J Mammal 70:740–753

Gannon WL, Sherwin RE, deCarvalho TN, O'Farrell MJ (2001) Pinnae and echolocation call differences between *Myotis californicus* and *M. ciliolabrum* (Chiroptera: Vespertilionidae). Acta Chiropterol 3(1):77–91

Gannon WL, O'Farrell MJ, Corben C, Bedrick EJ (2004) Call character lexicon and analysis of field recorded bat echolocation calls. In: Thomas J, Moss C, Vater M (eds) Echolocation in bats and dolphins. The University of Chicago Press, Chicago, pp 478–484

Garland EC, Castellote M, Berchok CL (2015) Beluga whale (*Delphinapterus leucas*) vocalizations and call classification from the eastern Beaufort Sea population. J Acoust Soc Am 137:3054–3067. https://doi.org/10.1121/1.4919338

Garland EC, Rendell L, Lilley MS, Poole MM, Allen J, Noad MJ (2017) The devil is in the detail: quantifying vocal variation in a complex, multi-levelled, and rapidly evolving display. J Acoust Soc Am 142(1): 460–472. https://doi.org/10.1121/1.4991320

Gavrilov AN, Parsons MJG (2014) A MATLAB tool for the characterization of recorded underwater sound (CHORUS). Acoust Aust 42(3):190–196

Gavrilov A, McCauley R, Gedamke J (2012) Steady inter and intra-annual decrease in the vocalization frequency of Antarctic blue whales. J Acoust Soc Am 131(6): 4476–4480. https://doi.org/10.1121/1.4707425

Gedamke J, Costa DP, Dunstan A (2001) Localization and visual verification of a complex minke whale vocalization. J Acoust Soc Am 109(6):3038–3047. https://doi.org/10.1121/1.1371763

Gemello R, Mana F (1991) A neural approach to speaker independent isolated word recognition in an uncontrolled environment. In: Proceedings of the International Neural Networks Conference, Paris 9–13 July 1990, vol 1. Kluwer Academic Publishers, Dordrecht, pp 83–86

Ghosh J, Deuser LM, Beck SD (1992) A neural network based hybrid system for detection, characterization, and classification of short-duration oceanic signals. IEEE J Ocean Eng 17:351–363. https://doi.org/10.1109/48.180304

Gill SA, Bierema AM-K (2013) On the meaning of alarm calls: a review of functional reference in avian alarm calling. Ethology 119:449–461. https://doi.org/10.1111/eth.12097

Gillespie D, Caillat M (2008) Statistical classification of odontocete clicks. Can Acoust 36:20–26

Gillespie D, Caillat M, Gordon J (2013) Automatic detection and classification of odontocete whistles. J Acoust Soc Am 134:2427–2437. https://doi.org/10.1121/1.4816555

Gingras G, Fitch WT (2013) A three-parameter model for classifying anurans into four genera based on advertisement calls. J Acoust Soc Am 133:547–559. https://doi.org/10.1121/1.4768878

Goëau H, Glotin H, Vellinga WP, Planqué R, Joly A (2016) LifeCLEF bird identification task 2016: the arrival of deep learning. CLEF 1609:440–449

Griffin DR, Webster FA, Michael CR (1960) The echolocation of flying insects by bats. Anim Behav 8:141–154

Guemeur Y, Elisseeff A, Paugam-Moisey H (2000) A new multi-class SVM based on a uniform convergence result. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium 4:183–188

Hamilton LJ, Cleary J (2010) Automatic discrimination of beaked whale clicks in noisy acoustic time series. In: OCEANS'10 IEEE Sydney, pp 1–5

Hammerschmidt K, Fischer J (1998) The vocal repertoire of Barbary macaques: a quantitative analysis of a graded signal system. Ethology 104(3):203–216. https://doi.org/10.1111/j.1439-0310.1998.tb00063.x

Hammerschmidt K, Reisinger E, Westekemper K, Ehrenreich L, Strenzke N, Fischer J (2012) Mice do not require auditory input for the normal development of their ultrasonic vocalizations. BMC Neurosci 13:40

Harland E (2008) Processing the workshop datasets using the TRUD algorithm. Can Acoust 36:27–33

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proc IEEE Conf Comput Vis Pattern Recogn 2016:770–778

Helweg DA, Cato ADH, Jenkins PF, Garrigue D, McCauley RD (1998) Geographic variation in South Pacific humpback whale songs. Behaviour 135:1–27

Herr, A, Klomp, NL, Atkinson, JS (1997) Identification of bat echolocation calls using decision tree classification system Complexity International. https://www.researchgate.net/publication/293134471_Identification_of_bat_echolocation_calls_using_a_decision_tree_classification_system. Accessed 17 July 2017

Himawan I, Towsey M, Law B, Roe P (2018). Deep learning techniques for Koala Activity detection. In: INTERSPEECH, pp. 2107–2111

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Holy TE, Guo Z (2005) Ultrasonic songs of male mice. PLoS One Biol 3(12):e386. https://doi.org/10.1371/journal.pbio.0030386

Horn AG, Falls JB (1996) Categorization and the design of signals: the case of song repertoires. In: Kroodsma DE, Miller EH (eds) Ecology and evolution of acoustic communication in birds. Comstock Publishing Associates, Ithaca, pp 121–135

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Edu Psychol 24: 417–441

Huang X, Acero A, Hon H-W (2001) Spoken language processing. Prentice Hall, Upper Saddle River, NJ

Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. Proc IEEE Conf Comput Vis Pattern Recogn 2017: 4700–4708

Ibrahim AK, Chérubin LM, Zhuang H, Schärer Umpierre MT, Dalgleish F, Erdol N, Ouyang B, Dalgleish A (2018) An approach for automatic classification of grouper vocalizations with passive acoustic monitoring. J Acoust Soc Am 143:666–676. https://doi.org/10.1121/1.5022281

Itakura F (1975) Minimum prediction residual principle applied to speech recognition. IEEE Trans Acoust Speech Sig Process 23:57–72

Jacobson EK, Yack TM, Barlow J (2013) Evaluation of an automated acoustic beaked whale detection algorithm

using multiple validation and assessment methods. In: NOAA Technical Memorandum NOAA-TM-NMFS-SWFSC-509

Jaitly N, Hinton GE (2013) Vocal tract length perturbation (VTLP) improves speech recognition. In: Proceedings of ICML Workshop on Deep Learning for Audio, Speech and Language, vol 117

Janik VM (1999) Pitfalls in the categorization of behavior: a comparison of dolphin whistle classification methods. Anim Behav 57:133–143. https://doi.org/10.1006/anbe.1998.0923

Jarvis S, Dimarzio N, Morrissey R, Moretti D (2006) Automated classification of beaked whales and other small odontocetes in the Tongue of the Ocean, Bahamas. Oceans 2006:1–6. https://doi.org/10.1109/OCEANS.2006.307124

Jiang JJ, Bu LR, Duan FJ, Wang XQ, Liu W, Sun ZB, Li CY (2019) Whistle detection and classification for whales based on convolutional neural networks. Appl Acoust 150:169–178. https://doi.org/10.1016/j.apacoust.2019.02.007

Kandia V, Stylianou Y (2006) Detection of sperm whale clicks based on the Teager–Kaiser energy operator. Appl Acoust 67(11):1144–1163. https://doi.org/10.1016/j.apacoust.2006.05.007

Karlsen JD, Bisther A, Lyndersen C, Haug T, Kovacs KM (2002) Summer vocalizations of adult male white whales (*Delphinapterus leucas*) in Svalbard, Norway. Polar Biol 25:808–817. https://doi.org/10.1007/s00300-002-0415-6

Keen S, Ross JC, Griffiths ET, Lanzone M, Farnsworth A (2014) A comparison of similarity-based approaches in the classification of flight calls of four species of North American wood-warblers (Parulidae). Ecol Inf 21:25–33. https://doi.org/10.1016/j.ecoinf.2014.01.001

Keighley MV, Langmore NE, Zdenek CN, Heinsohn R (2017) Geographic variation in the vocalizations of Australian palm cockatoos (*Probosciger aterrimus*). Bioacoustics 26(1):91–108. https://doi.org/10.1080/09524622.2016.1201778

Kershenbaum A, Blumstein DT, Roch MA, Akcay C, Backus G, Bee MA, Bohn K, Cao Y, Carter G, Cäsar C, Coen M, DeRuiter SL, Doyle L, Edelman S, Ferrer-i-Cancho R, Freeberg TM, Garland EC, Gustison M, Harley HE, Huetz C, Hughes M, Bruno JH, Ilany A, Jin DZ, Johnson M, Ju C, Karnowski J, Lohr B, Manser MB, McCowan B, Mercado E, Narins PM, Piel A, Rice M, Salmi R, Sasahara K, Sayigh L, Shiu Y, Taylor C, Vallejo EE, Waller S, Zamora-Gutierrez V (2016) Acoustic sequences in non-human animals: a tutorial review and prospectus. Biol Rev 91:13–52

Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114

Klinck H, Mellinger DK (2011) The energy ratio mapping algorithm: a tool to improve the energy-based detection of odontocete echolocation clicks. J Acoust Soc Am 129(4):1807–1812. https://doi.org/10.1121/1.3531924

Ko T, Peddinti V, Povey D, Khudanpur S (2015) Audio augmentation for speech recognition. In: Sixteenth Annual Conference of the International Speech Communication Association

Kogan J, Margoliash D (1998) Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study. J Acoust Soc Am 103:2185–2196. https://doi.org/10.1121/1.421364

Kollmorgen S, Hahnloser RH, Mante V (2020) Nearest neighbours reveal fast and slow components of motor learning. Nature 577(7791):526–530. https://doi.org/10.1038/s41586-019-1892-x

Kondo N, Watanabe S (2009) Contact calls: information and social function. Jpn Psych Res 51:197–208. https://doi.org/10.1111/j.1468-5884.2009.00399.x

Koren L, Geffen E (2009) Complex call in male rock hyrax (*Procavia capensis*): a multi-information distributing channel. Behav Ecol Sociobiol 63(4):581–590. https://doi.org/10.1007/s00265-008-0693-2

Koren L, Geffen E (2011) Individual identity is communicated through multiple pathways in male rock hyrax (*Procavia capensis*) songs. Behav Ecol Sociobiol 65(4):675–684. https://doi.org/10.1007/s00265-010-1069-y

Koren L, Mokady O, Geffen E (2008) Social status and cortisol levels in singing rock hyraxes. Horm Behav 54:212–216

Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90

Kruskal J, Sankoff D (1983) An anthology of algorithms and concepts for sequence comparison. In: Sankoff D, Kruskal J (eds) Time warps, string edits and macromolecules: the theory and practice of string comparison. Addison-Wesley, Reading, MA, pp 265–310

Lammers MO, Au WWL, Herzing DL (2003) The broadband social acoustic signaling behavior of spinner and spotted dolphins. J Acoust Soc Am 114:1629–1639. https://doi.org/10.1121/1.1596173

Law BS, Reinhold L, Pennay M (2002) Geographic variation in the echolocation sounds of *Vespadelus* spp. (Vespertilionidae) from New South Wales and Queensland, Australia. Acta Chiropt 4:201–215. https://doi.org/10.3161/001.004.0208

Le Boeuf BJ, Peterson RS (1969) Dialects in elephant seals. Science 166(3913):1654–1656. https://doi.org/10.1126/science.166.3913.1654

Leblanc E, Bahoura M, Simard Y (2008) Comparison of automatic classification methods for beluga whale vocalizations. J Acoust Soc Am 123:3772

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989a) Backpropagation applied to handwritten zip code recognition. Neural Comput 1(4):541–551. https://doi.org/10.1162/neco.1989.1.4.541

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989b) Handwritten digit recognition with a back-propagation network. In:

Proceedings of the 2nd International Conference on Neural Information Processing Systems, pp 396–404
LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791
LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539
Lee C-H, Hsu S-B, Shih J-L, Chou C-H (2013) Continuous birdsong recognition using Gaussian mixture modeling of image shape features. IEEE Trans Multimedia 15:454–464. https://doi.org/10.1109/TMM.2012.2229969
Leonard ML, Horn AG (2001) Begging calls and parental feeding decisions in tree swallows (*Tachycineta bicolor*). Behav Ecol Sociobiol 49:170–175. https://doi.org/10.1007/s002650000290
Levinson S (1985) Structural methods in automatic speech recognition. Proc IEEE 73:1625–1648. https://doi.org/10.1109/PROC.1985.13344
Li Z, Tang S, Yan S (2002) Multi-class SVM classifier based on pair wise coupling. In: Proceedings of the First International Workshop, SVM 2002, Niagara Falls, Canada, p 321
Liaw A, Wiener M (2002) Classification and regression by Random Forest. R News 2:18–22
Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y (2017) Efficient algorithms for t-distributed stochastic neighborhood embedding. arXiv preprint arXiv:1712.09005
Lippman R (1989) Pattern classification using neural networks. IEEE Commun Mag 1989:47–64
Luo W, Yang W, Zhang Y (2019) Convolutional neural network for detecting odontocete echolocation clicks. J Acoust Soc Am 145(1):EL7–EL12. https://doi.org/10.1121/1.5085647
Maaten LV (2014) Accelerating t-SNE using tree-based algorithms. J Mach Learn Res 15(1):3221–3245
Maaten LV, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
Mac Aodha O, Gibb R, Barlow KE, Browning E, Firman M, Freeman R, Harder B, Kinsey L, Mead GR, Newson SE, Pandourski I (2018) Bat detective—deep learning tools for bat acoustic signal detection. PLoS Comput Biol 14(3):e1005995. https://doi.org/10.1371/journal.pcbi.1005995
Madhusudhana S, Gavrilov AN, Erbe C (2015) Automatic detection of echolocation clicks based on a Gabor model of their waveform. J Acoust Soc Am 137(6):3077–3086. https://doi.org/10.1121/1.4921609
Madhusudhana S, Symes LB, Klinck H (2019) A deep convolutional neural network based classifier for passive acoustic monitoring of neotropical katydids. J Acoust Soc Am 146(4):2982–2982. https://doi.org/10.1121/1.5137323
Madhusudhana S, Murray A, Erbe C (2020) Automatic detectors for low-frequency vocalizations of Omura's whales, *Balaenoptera omurai*: a performance comparison. J Acoust Soc Am 147(5):3078–3090. https://doi.org/10.1121/10.0001108
Madhusudhana S, Shiu Y, Klinck H, Fleishman E, Liu X, Nosal EM, Helble T, Cholewiak D, Gillespie D, Širović A, Roch MA (2021) Improve automatic detection of animal call sequences with temporal context. J R Soc Interface 18:20210297. https://doi.org/10.1098/rsif.2021.0297
Madsen PT, Surlykke A (2013) Functional convergence in bat and toothed whale biosonars. Physiology 28(5):276–283. https://doi.org/10.1152/physiol.00008.2013
Makhoul J, Schwarz R (1995) State of the art in continuous speech recognition. Proc Nat Acad Sci USA 92:9956–9963. https://doi.org/10.1073/pnas.92.22.9956
Malfante M, Mohammed O, Gervaise C, Dalla Mura M, Mars JI (2018) Use of deep features for the automatic classification of fish sounds. In: 2018 OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO), pp 1–5. https://doi.org/10.1109/OCEANSKOBE.2018.8559276
Mankin RW, Smith T, Tropp JM, Atkinson EB, Young DY (2008) Detection of *Anoplophora glabripennis* (Coleoptera: Cerambycidae) larvae in different host trees and tissues by automated analysis of sound-impulse frequency and temporal patterns. J Econ Entomol 101(3):838–849. https://doi.org/10.1093/jee/101.3.838
Marler P (2004) Bird calls: a cornucopia for communication. In: Marler P, Slabbekoorn H (eds) Nature's music: the science of birdsong. Elsevier, Amsterdam, pp 132–177
Martindale S (1980a) On the multivariate analysis of avian vocalizations. J Theor Biol 83:107–110. https://doi.org/10.1016/0022-5193(80)90374-4
Martindale S (1980b) A numeric approach to the analysis of solitary vireo songs. Condor 82:199–211. https://doi.org/10.2307/1367478
Mazhar S, Ura T, Bahl R (2007) Vocalization based individual classification of humpback whales using support-vector-machine. Oceans 2007:1–9. https://doi.org/10.1109/OCEANS.2007.4449356
McDonald MA, Mesnick SL, Hildebrand JA (2006) Biogeographic characterisation of blue whale song worldwide: using song to identify populations. J Cetacean Res Manag 8(1):55–65
McInnes L, Healy J, Melville J (2018) UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426
McLaughlin J, Josso N, Ioana C (2008) Detection and classification of sound types in the vocalizations of north-east pacific blue whales. J Acoust Soc Am 123:3102
McLister D, Stevens ED, Bogart JP (1995) Comparative contractile dynamics of calling and locomotor muscles in three hylid frogs. J Exp Biol 198(7):1527–1538. https://doi.org/10.1242/jeb.198.7.1527
Mellinger DK (2008) A neural network for classifying clicks of Blainville's beaked whales (*Mesoplodon densirostris*). Can Acoust 36:55–59
Mellinger DK, Bradbury JW (2007) Acoustic measurement of marine mammal sounds in noisy

environments. In: Proceedings of the 2nd International Conference on Underwater Acoustic Measurements: Technologies and Results, Heraklion, Greece, 25–29 June 2007

Mellinger DK, Clark CW (2000) Recognizing transient low-frequency whale sounds by spectrogram correlation. J Acoust Soc Am 107(6):3518–3529. https://doi.org/10.1121/1.429434

Mellinger DK, Martin SW, Morrissey RP, Thomas L, Yosco JJ (2011) A method for detecting whistles, moans and other frequency contour sounds. J Acoust Soc Am 129:4055–4061. https://doi.org/10.1121/1.3531926

Mendelson TC, Shaw KL (2003) Rapid speciation in an arthropod. Nature 433:375–376. https://doi.org/10.1038/433375a

Mitani JC, Hasegawa T, Groslouis J, Marler P, Byrne R (1992) Dialects in wild chimpanzees. Am J Primatol 27:233–243

Møhl B, Wahlberg M, Madsen PT, Heerford A, Lund A (2003) The monopulsed nature of sperm whale sonar clicks. J Acoust Soc Am 114(2):1143–1154. https://doi.org/10.1121/1.1586258

Moon TK (1996) The expectation-maximization algorithm. IEEE Sig Process Mag 13:47–60. https://doi.org/10.1109/79.543975

Morrissey RP, Ward J, DiMarzio N, Jarvis S, Moretti DJ (2006) Passive acoustic detection and localization of sperm whales (*Physeter macrocephalus*) in the tongue of the ocean. Appl Acoust 67:1091–1105. https://doi.org/10.1016/j.apacoust.2006.05.014

Mouy X, Leary D, Martin B, Laurinolli M (2008) A comparison of methods for the automatic classification of marine mammal vocalizations in the Arctic. In: Proceedings of the PASSIVE'08 Workshop on New Trends for Environmental Monitoring using Passive Systems, Hyeres, France, 14–17 October 2008

Murray SO, Mercado E, Roitblat HL (1998) Characterizing the graded structure of false killer whale (*Pseudorca crassidens*) vocalizations. J Acoust Soc Am 104:1679–1687. https://doi.org/10.1121/1.424380

Myers C, Rabiner LR, Rosenberg AE (1980) Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. IEEE Trans Acoust Speech Sig Process 28:623–635. https://doi.org/10.1109/TASSP.1980.1163491

Nagy CM, Rockwell RF (2012) Identification of individual eastern screech-owls (*Megascops asio)* via vocalization analysis. Bioacoustics 21:127–140. https://doi.org/10.1080/09524622.2011.651829

Narins PM, Feng AS, Fay RR (eds) (2006) Hearing and sound communication in amphibians. Springer, New York

Noad MJ, Cato DH, Bryden MM, Jenner MN, Jenner KCS (2000) Cultural revolution in whale songs. Nature 408:537. https://doi.org/10.1038/35046199

Noda JJ, Travieso CM, Sánchez-Rodríguez D (2016) Automatic taxonomic classification of fish based on their acoustic signals. Appl Sci 6(12):443. https://doi.org/10.3390/app6120443

O'Farrell MJ, Miller BW, Gannon WL (1999) Qualitative identification of free-flying bats using Anabat detector. J Mammal 80:11–23

Oh J, Laubach M, Luczak A (2003) Estimating neuronal variable importance with random forest. Proc IEEE Bioeng Conf:33–34. https://doi.org/10.1109/NEBC.2003.1215978

Oleson EM, Širović A, Bayless AR, Hildebrand JA (2014) Synchronous seasonal change in fin whale song in the North Pacific. PLoS One 9(12):e115678. https://doi.org/10.1371/journal.pone.0115678

Oswald JN, Barlow J, Norris TF (2003) Acoustic identification of nine delphinid species in the eastern tropical Pacific Ocean. Mar Mamm Sci 19:20–37. https://doi.org/10.1111/j.1748-7692.2003.tb01090.x

Oswald JN, Rankin S, Barlow J, Lammers MO (2007) A tool for real-time acoustic species identification of delphinid whistles. J Acoust Soc Am 122:587–595. https://doi.org/10.1121/1.2743157

Oswald JN, Au WWL, Duennebier F (2011) Minke whale (*Balaenoptera acutorostrata*) boings detected at the Station ALOHA cabled observatory. J Acoust Soc Am 129:3353–3360. https://doi.org/10.1121/1.3575555

Oswald JN, Rankin S, Barlow J, Oswald M (2013) Real-time odontocete call classification algorithm: software for species identification of delphinid whistles. In: Adam O, Samaran F (eds) Detection, classification and localization of marine mammals using passive acoustics, 2003-2013: 10 years of international research. DIRAC NGO, Paris, France

Oswald JN, Walmsley SF, Casey C, Fregosi S, Southall B, Janik VM (2021) Species information in whistle frequency modulation patterns of common dolphins. Philos Trans R Soc B 376:20210046. https://doi.org/10.1098/rstb.2021.0046

Ou H, Au WWL, Oswald JN (2012) A non-spectrogram-correlation method of automatically detecting minke whale boings. J Acoust Soc Am 132:EL317–EL322

Ouattara K, Lemasson A, Zuberbunter K (2009) Campbell's monkeys concatenate vocalizations into context-specific call sequences. Proc Natl Acad Sci USA 106(51):22026

Papale E, Azzolin M, Cascao I, Gannier A, Lammers MO, Martin VM, Oswald JN, Perez-Gil M, Prieto R, Silva MA, Giacoma C (2013) Geographic variability in the acoustic parameters of striped dolphin's (*Stenella coeruleoalba*) whistles. J Acoust Soc Am 133:1126–1134. https://doi.org/10.1121/1.4774274

Papale E, Azzolin M, Cascao I, Gannier A, Lammers MO, Martin VM, Oswald J, Perez-Gil M, Prieto R, Silva MA, Giacoma C (2014) Macro- and micro- geographic variation of short-beaked common dolphin's whistles in the Mediterranean Sea and Atlantic Ocean. Ethol Ecol Evol 26:392–404. https://doi.org/10.1080/03949370.2013.851122

Park DS, Chan W, Zhang Y, Chiu C, Zoph B, Cubuk ED, Le QV (2019) SpecAugment: a simple data

augmentation method for automatic speech recognition. Proc Interspeech 2019:2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

Parsons S, Boonman AM, Obrist MK (2000) Advantages and disadvantages of techniques for transforming and analyzing chiropteran echolocation calls. J Mammal 81:927–938. https://doi.org/10.1644/1545-1542(2000)081<0927:AADOTF>2.0.CO;2

Payne K, Payne R (1985) Large scale changes over 19 years in songs of humpback whales in Bermuda. Z Tierpsychol 68:89–114. https://doi.org/10.1111/j.1439-0310.1985.tb00118.x

Picone JW (1993) Signal modeling techniques in speech recognition. Proc IEEE 81:1215–1247. https://doi.org/10.1109/5.237532

Placer J, Slobodchikoff CN (2000) A fuzzy-neural system for identification of species-specific alarm sounds of Gunnison's prairie dogs. Behav Process 52:1–9. https://doi.org/10.1016/S0376-6357(00)00105-4

Potter JR, Mellinger DK, Clark CW (1994) Marine mammal sound discrimination using artificial neural networks. J Acoust Soc Am 96:1255–1262. https://doi.org/10.1121/1.410274

Pozzi L, Gamba M, Giacoma C (2010) The use of Artificial Neural Networks to classify primate vocalizations: a pilot study on black lemurs. Am J Primatol 72(4):337–348. https://doi.org/10.1002/ajp.20786

Pröhl H, Koshy RA, Mueller U, Rand AS, Ryan MJ (2006) Geographic variation of genetic and behavioral traits in northern and southern Túngara frogs. Evol 60:1669–1679. https://doi.org/10.1111/j.0014-3820.2006.tb00511.x

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–285

Rabiner LR, Juang BH (1986) An introduction to Hidden Markov Models. IEEE ASSP Mag 1986:4–16

Rabiner LR, Levinson S, Sondhi M (1983) On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition. Bell Syst Tech J 62:1075–1106. https://doi.org/10.1002/j.1538-7305.1983.tb03115.x

Rabiner LR, Juang B, Lee C (1996) An overview of automatic speech recognition. In: Lee C, Soong F, Paliwal K (eds) Automatic speech and speaker recognition. Kluwer Academic, New York, pp 1–30

Rankin S, Barlow J (2005) Source of the North Pacific 'boing' sound attributed to minke whales. J Acoust Soc Am 118(5):3346–3351. https://doi.org/10.1121/1.2046747

Rankin S, Ljungblad D, Clark CW, Kato H (2005) Vocalisations of Antarctic blue whales, *Balaenoptera musculus intermedia*, recorded during the 2001/2002 and 2002/2003 IWC/SOWER circumpolar cruises, Area V, Antarctica. J Cet Res Manag 7(1):13–20

Rankin S, Archer F, Keating JL, Oswald JN, Oswald M, Curtis A, Barlow J (2016) Acoustic classification of dolphins in the California Current using whistles, clicks and burst-pulses. Mar Mamm Sci 33:520–540. https://doi.org/10.1111/mms.12381

Reby D, André-Obrecht R, Galinier A, Farinas J, Cargnelutti B (2006) Cepstral coefficients and hidden Markov models reveal idiosyncratic voice characteristics in red deer (*Cervus elaphus*) stags. J Acoust Soc Am 120:4080–4089. https://doi.org/10.1121/1.2358006

Recalde-Salas A, Salgado Kent CP, Parsons MJG, Marley SA, McCauley RD (2014) Non-song vocalizations of pygmy blue whales in Geographe Bay, Western Australia. J Acoust Soc Am 135(5):EL213–EL218. https://doi.org/10.1121/1.4871581

Recalde-Salas A, Erbe C, Salgado Kent C, Parsons M (2020) Non-song vocalizations of humpback whales in Western Australia. Front Mar Sci 7:141. https://doi.org/10.3389/fmars.2020.00141

Rickwood P, Taylor A (2008) Methods for automatically analyzing humpback song units. J Acoust Soc Am 123:1763–1772. https://doi.org/10.1121/1.2836748

Risch D, Gales NJ, Gedamke J, Kindermann L, Nowacek DP, Read AJ, Siebert U, Van Opzeeland IC, Van Parijs SM, Friedlander AS (2014) Mysterious bio-duck sound attributed to the Antarctic minke whale (*Balaenoptera bonaerensis*). Biol Lett 10:20140175. https://doi.org/10.1098/rsbl.2014.0175

Roch MA, Soldevilla MS, Burtenshaw JC, Henderson EE, Hildebrand JA (2007) Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California. J Acoust Soc Am 121:1737–1748. https://doi.org/10.1121/1.2400663

Roch MA, Soldevilla MS, Hoenigman R, Wiggins SM, Hildebrand JA (2008) Comparison of machine-learning techniques for the classification of echolocation clicks from three species of odontocetes. Can Acoust 36:41–47

Roch MA, Brandes TS, Patel B, Barkley Y, Baumann-Pickering S, Soldevilla MS (2011) Automated extraction of odontocete whistle contours. J Acoust Soc Am 130:2212–2223. https://doi.org/10.1121/1.3624821

Rocha HS, Ferreira LS, Paula BC, Rodrigues HG, Sousa-Lima RS (2015) An evaluation of manual and automated methods for detecting sounds of mane wolves (*Chrysocyon brachyurus* Illiger 1815). Bioacoustics 24:185–198. https://doi.org/10.1080/09524622.2015.1019361

Roitblat HL, Moore PWB, Nachtigall PE, Penner RH, Au WWL (1989) Natural echolocation with an artificial neural network. Int J Neural Syst 1:239–247

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 65:386–408. https://doi.org/10.1037/h0042519

Ross JC, Allen PE (2014) Random forest for improved analysis efficiency in passive acoustic monitoring. Ecol Inform 21:34–39. https://doi.org/10.1016/j.ecoinf.2013.12.002

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536. https://doi.org/10.1038/323533a0

Russo D, Mucedda M, Bello M, Biscardi S, Pidinchedda E, Jones G (2007) Divergent echolocation sound frequencies in insular rhinolophids (Chiroptera): a case of character displacement? J Bioeng 34:2129–2138. https://doi.org/10.1111/j.1365-2699.2007.01762.x

Sainburg T, Theilman B, Thielk M, Gentner TQ (2019) Parallels in the sequential organization of birdsong and human speech. Nat Commun 10:3636. https://doi.org/10.1038/s41467-019-11605-y

Sakoe H, Chiba S (1978) Dynamic programming optimization for spoken word recognition. IEEE Trans Acoust Speech Sig Process 26:43–49. https://doi.org/10.1109/TASSP.1978.1163055

Schassburger RM (1993) Vocal communication in the timber wolf, *Canis lupus*, Linnaeus: structure, motivation, and ontogeny. Parey Scientific Publication, New York

Schon PC, Puppe B, Manteauffel G (2001) Linear prediction coding analysis and self-organizing feature map as tools to classify stress sounds of domestic pigs (*Sus scrofa*). J Acoust Soc Am 110:1425–1431. https://doi.org/10.1121/1.1388003

Sethi SS, Jones NS, Fulcher BD, Picinali L, Clink DJ, Klinck H, Orme CD, Wrege PH, Ewers RM (2020) Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. Proc Natl Acad Sci 117(29):17049–17055. https://doi.org/10.1073/pnas.2004702117

Shannon CE, Weaver W (1998) The mathematical theory of communication. University of Illinois Press, Champaign

Shiu Y, Palmer KJ, Roch MA, Fleishman E, Liu X, Nosal EM, Helble T, Cholewiak D, Gillespie D, Klinck H (2020) Deep neural networks for automated detection of marine mammal species. Sci Rep 10(1):1–12. https://doi.org/10.1038/s41598-020-57549-y

Sibley DA (2000) The Sibley field guide to birds. Knopf, New York

Simmons JA, Wever EG, Pylka JM (1971) Periodical cicada: sound production and hearing. Science 171(3967):212–213. https://doi.org/10.1126/science.171.3967.212

Širović A (2016) Variability in the performance of the spectrogram correlation detector for north-east Pacific blue whale calls. Bioacoustics 25(2):145–160. https://doi.org/10.1080/09524622.2015.1124248

Širović A, Cutter GR, Butler JL, Demer DA (2009) Rockfish sounds and their potential use for population monitoring in the Southern California Bight. ICES J Mar Sci 66:981–990. https://doi.org/10.1093/icesjms/fsp064

Sjare B, Stirling I, Spencer C (2003) Seasonal and longer-term variability in the songs of Atlantic walruses breeding in the Canadian High Arctic. Aquat Mamm 29(2):297–318

Slobodchikoff CN, Ackers SH, Van Ert M (1998) Geographic variation in alarm calls of Gunnison's prairie dogs. J Mammal 79(4):1265–1272. https://doi.org/10.2307/1383018

Somervuo P, Härmä A, Fagerlund S (2006) Parametric representations of bird sounds for automatic species recognition. IEEE Trans Audio Speech Lang Process 14:2252–2263. https://doi.org/10.1109/TASL.2006.872624

Sparling DW, Williams JD (1978) Multivariate analysis of avian vocalizations. J Theor Biol 74:83–107. https://doi.org/10.1016/0022-5193(78)90291-6

Stafford KM, Fox CG, Clark DS (1998) Long-range acoustic detection and localization of blue whale sounds in the northeast Pacific Ocean. J Acoust Soc Am 104(6):3616–3625. https://doi.org/10.1121/1.423944

Stafford KM, Nieukirk SL, Fox CG (1999) Low-frequency whale sounds recorded on hydrophones moored in the eastern tropical Pacific. J Acoust Soc Am 106:3687–3698. https://doi.org/10.1121/1.428220

Stafford KM, Moore SE, Laidre KL, Heide-Jørgensen MP (2008) Bowhead whale springtime song off West Greenland. J Acoust Soc Am 124(5):3315–3323. https://doi.org/10.1121/1.2980443

Starnberger I, Preininger D, Hödl W (2014) The anuran vocal sac: a tool for multimodal signalling. Anim Behav 97:281–288. https://doi.org/10.1016/j.anbehav.2014.07.027

Stoeger AS, Heilmann G, Zeppelzauer M, Ganswindt A, Hensman S, Charlton BD (2012) Visualizing sound emission of elephant vocalizations: evidence for two rumble production types. PLoS One 7:1–8. https://doi.org/10.1371/journal.pone.0048907

Stowell D, Wood M, Stylianou Y, Glotin H (2016). Bird detection in audio: a survey and a challenge. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), pp 1–6. https://doi.org/10.1109/MLSP.2016.7738875

Sturtivant C, Datta S (1997) Automatic dolphin whistle detection, extraction, encoding, and classification. Proc Inst Acoust 19:259–266

Suzuki R, Buck J, Tyack P (2006) Information entropy of humpback whale songs. J Acoust Soc Am 119:1849–1866. https://doi.org/10.1121/1.2161827

Swets JA, Dawes RM, Monahan J (2000) Better decisions through science. Sci Am 283:82–87

Takahashi N, Kashino M, Hironaka N (2010) Structure of rat ultrasonic vocalizations and its relevance to behavior. PLoS One 5(11):e14115. https://doi.org/10.1371/journal.pone.0014115

Tan M, McDonald K (2017) Bird sounds | Experiments with Google [online]. https://experiments.withgoogle.com/bird-sounds

Tchernichovski O, Nottebohm F, Ho CE, Pesaran B, Mitra PP (2000) A procedure for an automated measurement of song similarity. Anim Behav 59:1167–1176. https://doi.org/10.1006/anbe.1999.1416

Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323. https://doi.org/10.1126/science.290.5500.2319

Thomas JA, Golladay CL (1995) Analysis of underwater vocalizations of leopard seals (*Hydrurga leptonyx*). In: Kastelein RA, Thomas JA, Nachtigall PE (eds) Sensory systems of aquatic mammals. De Spil Publishers, Amsterdam, pp 201–221

Thomas M, Martin B, Kowarski K, Gaudet B, Matwin S (2019) Marine mammal species classification using convolutional neural networks and a novel acoustic representation. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp 290–305

Torrey L, Shavlik J (2010) Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI Global, New York, pp 242–264

Trawicki MB, Johnson MT, Osiejuk TS (2005) Automatic song-type classification and speaker identification of Norwegian ortolan bunting. IEEE Int Conf Mach Learn Sig Process (MLSP) 2005:277–282. https://doi.org/10.1109/MLSP.2005.1532913

Trifa VM, Kirschel ANG, Taylor CE (2008) Automated species recognition of antbirds in a Mexican rainforest using hidden Markov Models. J Acoust Soc Am 123:2424–2431. https://doi.org/10.1121/1.2839017

Valente D, Wang H, Andrews P, Mitra PP, Saar S, Tchernichovski O, Golani I, Benjamini Y (2007) Characterizing animal behavior through audio and video signal processing. IEEE Multimedia 14:32–41. https://doi.org/10.1109/MMUL.2007.71

Van Allen E, Menon MM, Dicaprio N (1990) A modular architecture for object recognition using neural networks. In: Proceedings of International Neural Networks Conference, Paris, vol 1, pp 35–379, 13 July 1990. Kluwer Academic Publishers, Dordrecht

Vapnik VN (1998) Statistical learning theory. Wiley, New York

Venter PJ, Hanekom JJ (2010) Automatic detection of African elephant (*Loxodonta africana*) infrasonic vocalizations from recordings. Biosyst Eng 106:286–294. https://doi.org/10.1016/j.biosystemseng.2010.04.001

Von Muggenthaler E, Reinhart P, Lympany B, Craft RB (2003) Songlike vocalizations from the Sumatran rhinoceros (*Dicerorhinus sumatrensis*). Acoust Res Lett 4(3):83–88. https://doi.org/10.1121/1.1588271

Waibel A, Hanazawa T, Hinton G, Shikano K, Lang KL (1989) Phoneme recognition using time-delay neural networks. IEEE Trans Acoust Speech Signal Proc 37:328–339. https://doi.org/10.1109/29.21701

Ward J, Morrissey R, Moretti D, DiMarzio N, Jarvis S, Johnson M, Tyack PL, White C (2008) Passive acoustic detection and localization of *Mesoplodon densirostris* (Blainville's beaked whale) vocalizations using distributed bottom-mounted hydrophones in conjunction with a digital tag (DTag) recording. Can Acoust 36:60–66

Ward R, Parnum I, Erbe C, Salgado-Kent CP (2016) Whistle characteristics of Indo-Pacific bottlenose dolphins (*Tursiops aduncus*) in the Fremantle Inner Harbour, Western Australia. Acoust Aust 44(1):159–169. https://doi.org/10.1007/s40857-015-0041-4

Ward R, Gavrilov AN, McCauley RD (2017) "Spot" call: A common sound from an unidentified great whale in Australian temperate waters. J Acoust Soc Am 142(2):EL231–EL236. https://doi.org/10.1121/1.4998608

Weisburn BA, Mitchell SG, Clark CW, Parks TW (1993) Isolating biological acoustic transient signals. Proc IEEE Int Conf Acoust Speech Sig Process 1:269–272. https://doi.org/10.1109/ICASSP.1993.319107

Wellard R, Erbe C, Fouda L, Blewitt M (2015) Vocalisations of killer whales (*Orcinus orca*) in the Bremer Canyon, Western Australia. PLoS One 10(9):e0136535. https://doi.org/10.1371/journal.pone.0136535

Wells KD (2007) The ecology and behaviour of amphibians. University of Chicago Press, Chicago, IL

Wich SA, Schel AM, De Vries H (2008) Geographic variation in Thomas langur (*Presbytis thomasi*) loud sounds. Am J Primatol 70:566–574. https://doi.org/10.1002/ajp.20527

Winn HE, Winn LK (1978) The song of the humpback whale *Megaptera novaeangliae* in the West Indies. Mar Biol 47:97–114. https://doi.org/10.1007/BF00395631

Wood JD, McCowan B, Langbauer WR, Viljoen JJ, Hart LA (2005) Classification of African elephant *Loxodonta africana* rumbles using acoustic parameters and cluster analysis. Bioacoustics 15:143–161. https://doi.org/10.1080/09524622.2005.9753544

Yamamoto O, Moore B, Brand L (2001) Variation in the bark sound of the red squirrel (*Tamiasciurus hudsonicus*). West N Am Nat 61:395–402

Yang X-J, Lei F-M, Wang G, Jesse AJ (2007) Syllable sharing and inter-individual syllable variation in Anna's hummingbird *Calypte anna* songs, in San Francisco, California. Folia Zool 56:307–318

Yoshino H, Armstrong KN, Izawa M, Yokoyama J, Kawata M (2008) Genetic and acoustic population structuring in the Okinawa least horseshoe bat: are intercolony acoustic differences maintained by vertical maternal transmission? Mol Ecol 17:4978–4991. https://doi.org/10.1111/j.1365-294X.2008.03975.x

Zar JH (2009) Biostatistical analysis, 5th edn. Pearson, New York, p 960

Zeppelzauer M, Hensman S, Stoeger AS (2015) Towards an automated acoustic detection system for free-ranging elephants. Bioacoustics 24:13–29. https://doi.org/10.1080/09524622.2014.906321

Zhang YJ, Huang JF, Gong N, Ling ZH, Hu Y (2018) Automatic detection and classification of marmoset vocalizations using deep and recurrent neural

networks. J Acoust Soc Am 144(1):478–487. https://doi.org/10.1121/1.5047743

Zhong M, LeBien J, Campos-Cerqueira M, Dodhia R, Ferres JL, Velev JP, Aide TM (2020) Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. Appl Acoust 166:107375. https://doi.org/10.1016/j.apacoust.2020.107375

Zuberbuhler K, Jenny D, Bshary R (1999) The predator deterrence function of primate alarm calls. Ethology 105:477–490. https://doi.org/10.1046/j.1439-0310.1999.00396.x