Chapter 15
Signal Processing

David K. Mellinger, Marie A. Roch, Eva-Marie Nosal, and Holger Klinck

Abstract We examine some methods commonly used for analyzing marine bioacoustic recordings. Filtering techniques are used to prevent aliasing, to remove certain types of noise, to flatten the spectrum of ocean noise before recording, and so on. Filter design necessarily requires making choices that affect trade-offs among various desirable filter properties. Detection and classification are used for analyzing large data sets. They often start with signal conditioning, which can adjust the spectrum, standardize signal level, and remove some types of noise. They proceed by calculating numerical acoustic features and using them to decide whether a given sound is present (detection) or to choose which of several categories a vocalization belongs to (classification). A variety of methods for detection and classification are briefly described, with the choice depending both on the nature of the sound(s) and the noise as well as on the task to be solved. Detectors operate in the time domain or on a time–frequency representation, with different ones appropriate for different call types. Classifiers are characterized as either generative or discriminative, as parametric or nonparametric, and as supervised or non-supervised. Performance of detection and classification can be evaluated in several ways, including receiver operating characteristic curves and precision/recall statistics. Localization of calling animals is usually performed using time differences of arrival of sounds at several hydrophones; a variety of methods are available, with the best choice depending on the characteristics of the sound and the acoustic environment. The most accurate localization methods use acoustic propagation modeling to estimate travel times. Several software packages are reviewed for filtering, detection, classification, and localization.

D.K. Mellinger (✉) • H. Klinck
Cooperative Institute for Marine Resources Studies, Hatfield Marine Science Center, Oregon State University, 2030 SE Marine Science Drive, Newport, OR 97365, USA
e-mail: David.Mellinger@oregonstate.edu

M.A. Roch
Department of Computer Science, San Diego State University, 5500 Campanile Dr., San Diego, CA 92182-7720, USA

E.-M. Nosal
School of Ocean and Earth Science and Technology, University of Hawaii at Manoa, 1680 East–west Road, Honolulu, HI 81622, USA

W.W.L. Au, M.O. Lammers (eds.), *Listening in the Ocean*, Modern Acoustics and Signal Processing, DOI 10.1007/978-1-4939-3176-7_15

15.1 Introduction

Marine animal sounds are captured using the systems covered in the preceding chapters of this book and ones similar to them. These systems use hydrophones to capture a sound signal—a representation of the sound pressure over time—and either make it available in real time or store it for later analysis. Analysis of biological and anthropogenic sounds has the potential to provide the kinds of information used in the previous chapters in this book—census information (presence/absence or counts), habitat usage, insights into behavior, and the effect of human activities on marine life. This chapter provides an introduction to the *signal processing* needed to accomplish these tasks. Throughout the chapter, it is the authors' intention to provide a qualitative description of common signal processing techniques along with references as to guide the reader interested in acquiring in-depth knowledge. The type of signal processing needed depends on the type of result desired. For instance, assessing the possibility of physiological harm to an animal (temporary or permanent deafness, tissue damage, etc.) requires knowing the sound spectrum received by the animal over time. This in turn may require signal processing techniques to localize the animal from its calls and to measure the sound spectrum over time. To study a species' distribution or movement, one can automatically detect, and sometimes localize, vocalizations from individuals of that species.

Here we review some of the most common signal processing tasks employed in marine bioacoustics. We assume that the reader is familiar with Fourier transforms and their properties, at least at a conceptual level. First is a section on filtering, which is commonly used in data acquisition, resampling of signals, and flattening of spectral responses. Following that is a description of automatic call detection and classification, reviewing the strengths and weaknesses of the more widely used methods, and then a discussion of localization techniques and applications. Final sections cover software widely used for marine bioacoustics as well as future directions for research.

15.2 Filtering

Filtering is commonly used in marine bioacoustics to alter the spectrum of a sound signal. A filter receives a sound signal as input, alters it in some manner, and emits the altered signal. For instance, a *low-pass filter* allows lower frequencies to pass through unimpeded but stops higher frequencies (Fig. 15.1). The frequency regions where sound is allowed through the filter is known as the pass band and frequencies that are attenuated are within the stop band. The point of transition between a pass band and a stop band is referred to as the cutoff frequency, or sometimes the break frequency or corner frequency. Conversely, a *high-pass filter* allows the high frequencies to pass through and stops the lower ones. A *band-pass filter* passes through only a selected range or band of frequencies, blocking frequencies above and below that range; it has two corner frequencies, one pass band, and two stop bands.

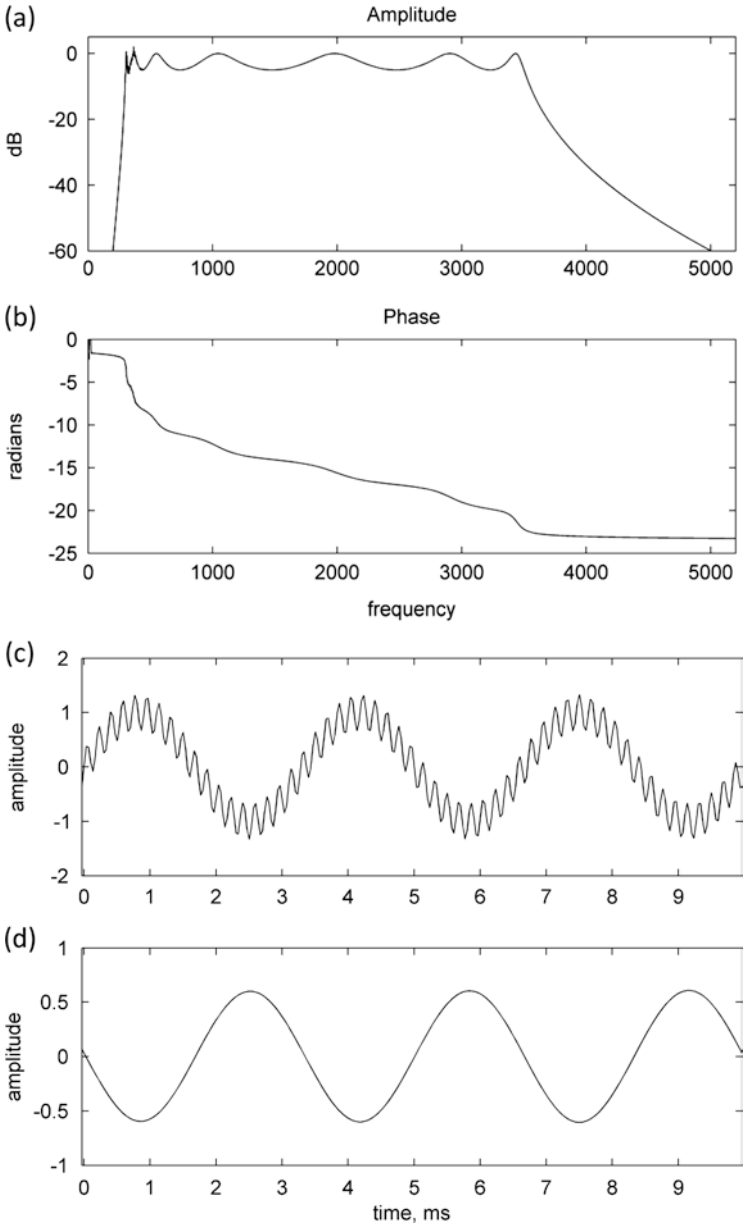

Fig. 15.1 Filter frequency response in (**a**) amplitude and (**b**) phase showing the pass band (300–3500 Hz) and stop bands (<50 and >5100 Hz). Note the ripple in the amplitude pass band, as well as the imperfect linearity of the phase in the pass band; a perfectly linear filter would have a *straight line* in this region. The phase in the stop bands is highly nonlinear, but this is relatively unimportant since there is very little signal energy at these frequencies. This is a 7th-order Chebyshev Type I IIR filter. (**c**) A signal with sinusoidal components at 300 and 5500 Hz, and (**d**) the same signal after filtering with this filter. The 5500 Hz component is removed, as it is in the stop band. Note that the phase of the 300 Hz signal is shifted by $-\pi$ radians (180°) as predicted by the phase plot (**b**) at 300 Hz

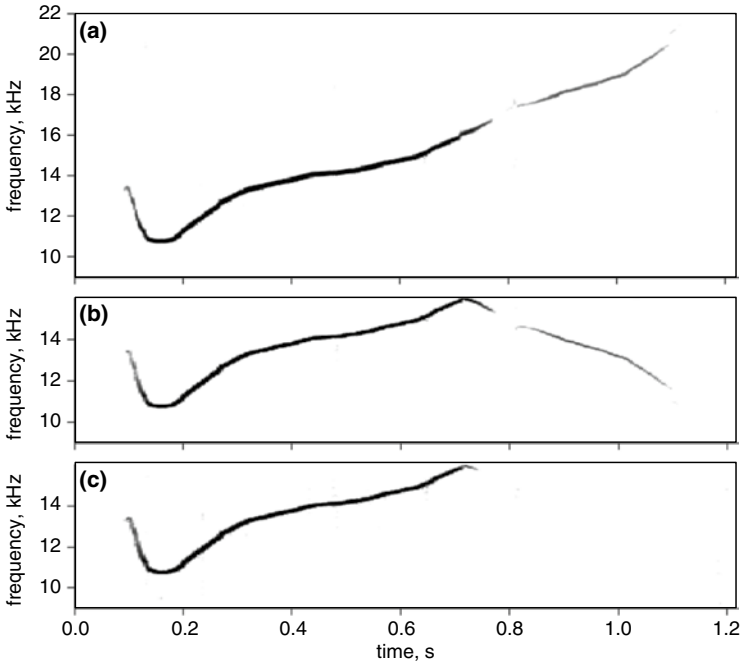

Fig. 15.2 Spectrograms showing an example of aliasing of a common dolphin whistle. (**a**) The whistle with a sample rate sufficiently high to capture it in its entirety. (**b**) The sound improperly resampled without low-pass filtering. At the Nyquist frequency of 16 kHz, the whistle appears to "reflect" to lower, incorrect frequencies. (**c**) The sound properly resampled, using filtering before resampling with a low-pass cutoff of 16 kHz. The *top part* of the dolphin whistle above the Nyquist frequency is absent, as it must be at this sample rate, and no longer appears at the wrong frequency

The most common use for filtering is to prevent aliasing. Aliasing occurs when sounds are present above the Nyquist frequency, which is defined as half the sampling rate. When such sounds are represented as a digital signal, they are indistinguishable from sounds below the Nyquist frequency—in other words, they appear *aliased* to that lower frequency (Fig. 15.2). Sound-playback equipment will play them as the lower frequency. To prevent this frequency shift, every digital acquisition system—every system for converting analog signals into digital samples—has an analog anti-alias low-pass filter to remove sounds above the Nyquist frequency.

Anti-alias low-pass filtering is also necessary when resampling a digital signal to a new sampling rate: Because the Nyquist frequency for the new sampling rate is different from that of the old rate, all sounds at frequencies above the new Nyquist rate must be removed from the signal before it is resampled at the new rate. To down-sample a signal to $1/k$ of its current sampling rate r, then, one must apply a low pass filter with a cutoff frequency of $r/(2k)$, then down-sample by selecting every kth sample of the filtered signal. To up-sample a signal to k times its current sampling rate r, one must insert $k-1$ zeroes after every sample to obtain a signal with the

desired sampling rate of *kr*, then apply a low-pass filter to the new signal with a cutoff frequency of *r*/2.

Another relatively common use of filters is to flatten the spectral response of hardware devices. A hydrophone, for instance, may capture some frequencies well but attenuate others somewhat. To correct this spectral shaping, a filter can be designed with the inverse of the hydrophone's spectral response, thus restoring the original spectrum of the sound signal.

Different filters have different properties, and it is helpful to understand the tradeoffs between these properties in choosing a type of filter.

- The most prominent property is the *frequency response* of a filter, which specifies how much gain or attenuation the filter causes at each frequency between 0 Hz and the Nyquist frequency. Often one desires a filter with a "rectangular" frequency response, such that all frequencies in the pass band have gain 1 (0 dB) and all other frequencies have gain 0 ($-\infty$ dB). Unfortunately, this is mathematically impossible for a finite filter, and all realizable filters are only an approximation to this ideal filter. Common ways in which a filter misses the ideal are (a) having *ripple* in the pass-band, such that the gain oscillates above and below 1; (b) having *transition region(s)* of some bandwidth in which the gain goes from 1 to nearly 0 or vice versa; often one wants this transition region to occupy only a narrow band of frequencies; (c) having the gain in the stop band be some number of decibels below the gain in the pass band, rather than the ideal gain of 0; attenuation of 60 dB in the stop band is often used to remove unwanted frequencies. Examples of these shortfalls can be seen in Fig. 15.1a. Generally speaking, all of these properties improve with increasing order of the filter (see below).
- A related value is the *phase response*, which specifies (in degrees or radians) how much each frequency is delayed as it passes through the filter. Identical delay across frequencies is also called *linear phase response*, since a constant delay time is the same as a phase change that increases linearly with frequency. Having a constant time delay can be important when detecting calls using templates, when analyzing call characteristics, or in any other application for which the shape of the call is important.
- The *order* (length) of the filter, usually denoted by *N*. The frequency response generally improves with increasing order, but at a price: The *computational cost* of a filter, which is important for real-time applications, is proportional to *N*. This is discussed in more detail below.
- The *response time* of a filter refers to the time it takes for a given sound on input to appear (filtered) at the output. Response time is also called *group delay*. For most filters, response time is also proportional to *N*, and for many filters it is equal to *N*/2 sample periods. Response time for some filters (IIR filters, described below) can vary with frequency.
- *Stability* is a factor for some filters. An unstable filter can, with certain inputs, have an output that increases toward infinity.

Generally speaking, one can improve the frequency and phase responses—make the pass band have less ripple, make the transition region narrower, or decrease the gain

in the stop band—by increasing the order of the filter. The drawback is that the computational cost rises, and usually the response time does as well.

An important distinction in digital filters is whether they are *infinite impulse response* (IIR) or *finite impulse response* (FIR). These are also called *recursive* and *non-recursive* filters, respectively. An IIR filter reuses one or more of its previous output values in computing the next output value (hence the name recursive), while an FIR filter does not. Because IIR filters have this feedback, they can be unstable. A more complete discussion of stability is available elsewhere (Oppenheim and Schafer 2009), but suffice it to say that one can test a digital filter for stability by providing it an impulse—a signal whose samples are all zero-valued except for a single 1 value—and checking whether the filter's output decays to 0 over time.

A digital filter consists essentially of two length $N+1$ vectors of filter coefficients, traditionally called $\vec{A}$ and $\vec{B}$, where N is the order of the filter. Many methods for designing digital filters are available, including IIR filter design methods known as Chebyshev types I and II, elliptical, Butterworth, and Bessel, and FIR methods called the window method and the frequency-sampling method. A more complete discussion of all these methods is available elsewhere (Oppenheim and Schafer 2009), but one can judge a given filter by examining its frequency and phase responses (Fig. 15.1) and considering its order.

The filter coefficients are used to implement the filter. In simplest form, a filter is implemented with

$$y[n] = (1/a_0)\begin{pmatrix} b_0 x[n] + b_1 x[n-1] + b_2 x[n-2] + \ldots \\ -a_1 y[n-1] - a_2 y[n-1] - \ldots \end{pmatrix} \tag{15.1}$$

where $x[n]$ is the input signal, $y[n]$ is the output signal, n is a time index (with smaller values in the past), and a_i and b_i are the filter coefficient vectors. When implementing a filter to operate on successive blocks of input data, care must be taken to preserve the data from the end of one block for the start of the next block to prevent a discontinuity in the output signals. For FIR filters, one must preserve the last N inputs $x[n]$; for IIR filters, one must preserve both these inputs and also the last N outputs $y[n]$. Equivalently, it is possible to apply FIR filters by preserving only the input samples: If the block length is m samples, with $m \gg N$, one can filter each block and then ignore the first N and last N samples of the result, keeping only the middle $m-2N$ samples. (Thus the start of each input block must be $m-2N$ samples after the start of the previous block in the input sample stream.) FIR filtering can also be implemented using a discrete Fourier transform (DFT) to perform the convolution represented by Eq. (15.1). This is computationally more efficient, sometimes dramatically so, for filters whose order is more than a handful of samples. and can be combined with FIR block processing as described above. See Oppenheim and Schafer (2009) for information on performing convolution using a DFT.

For an FIR filter, all A coefficients in Eq. (15.1) after a_0 are zero; a_0 is often 1, so it can be ignored as well. The computational cost of an FIR filter of order N is thus half that of an IIR filter of the same order. Also, because the filter's output depends on only the inputs x and not the outputs y, the filter is inherently stable; once the

most recent $N+1$ inputs in an impulse signal are all zero, the output of the FIR filter is necessarily zero as well. FIR filters can be (and usually are) designed to be symmetric, with the left half of the coefficients a mirror image of the right half, which implies that they have constant time delay, or linear phase (Oppenheim and Schafer 2009). The drawback of FIR filters is that they tend to have much higher order, and hence much higher computational cost and longer response time, for a given frequency response. For instance, an FIR low-pass filter of order 110 has approximately the same transition band width—i.e., its frequency response falls off just as fast above the cutoff frequency—as an IIR Chebyshev Type I filter of order 10. Although the computational cost of the FIR filter is half that of an equal-order IIR filter, this FIR filter still has 5.5 times the computational cost and response time of the comparable IIR filter.

Digital filters can be designed using several popularly available packages (such as in the Signal Processing Toolbox in MATLAB™), or via websites that allow one to enter the desired filter characteristics and then return the filter coefficient vectors. The packages also contain methods that implement Eq. (15.1)—that apply the filter to a block of input samples $x(n)$ and return the output samples $y(n)$, with provisions for preserving the filter's state between the end of one block and the start of the next. These routines are usually highly optimized to run quickly, using a DFT and employing multiple processors when possible. So if they are available, by all means use them.

15.3 Detection and Classification

Many applications of marine bioacoustics involve large-scale data sets—data sets collected from many hydrophones, or over long time periods, or both. Analyzing such data sets usually requires automated methods to find any animal vocalizations of interest. This process may be broken down into the separate steps of *detection*—finding potential sounds of interest in the recorded signal—and *classification*, assigning these sounds to categories. Detection methods usually operate on a continuous signal, making decisions at each time step about whether a sound of interest is present or not, while classification methods operate on short, discrete chunks of sound, typically ones roughly the duration of the calls under investigation, to assign them to one of several categories. Despite these differences, there is no firm distinction between detection and classification, and many techniques do some of both. Even a detection method as simple as finding any transient sound typically operates in a specific frequency band and with transients of a certain duration, characteristics that cause it to have some selectivity for—some classification of—the sounds it detects. Classification techniques are sometimes used in a two-way decision to choose whether a sound is of a desired call type or not, a task that is very much like detection. In addition, some techniques, like the template-matching methods discussed below, combine detection and classification into one step. Also, for the common case in which detection is followed by classification, the sounds found by the detector and thus presented to the classifier are very much dependent on the

characteristics of the detector, and training and testing of the two methods is closely intertwined. This section reviews some of the issues that arise in using detection and classification methods, and succeeding sections examine some of the widely used methods.

Detection and classification methods typically use one or more *features* of a sound to make decisions. A feature, also known as a *measurement*, *statistic*, or *observation*, is simply a quantity derived (extracted) from the sound by some algorithm. Examples include minimum frequency, duration, amplitude modulation, and entropy (e.g., Erbe and King 2008). There can be multiple algorithms for a given type of measurement; for instance, measuring duration is not simple for sounds that fade in amplitude at the end, and it can be done in several ways. Fristrup (1992) developed noise-robust methods for estimating features of animal sounds. Detection and classification systems often use several features calculated from each sound, in which case the features are grouped into a *feature vector* containing all of the desired values. (Confusingly, the feature *vector* is sometimes referred to as the feature or observation itself.) An *N*-element feature vector, corresponding to one call, or portion thereof, defines a single point in an *N*-dimensional space, and the implicit or explicit goal of many classifiers is to group as a single class those points that are near each other in this space. Using multiple feature vectors to represent a call is common when there is some form of evolution over time of the signal. An example of this can be seen in Deecke and Janik's (2006) work with dolphin whistles where measurements of the signal were produced every 10 ms.

Decision criteria. Detection and classification tasks use a *decision criterion* to decide the class, if any, to which a segment of audio belongs. Fig. 15.3a shows an

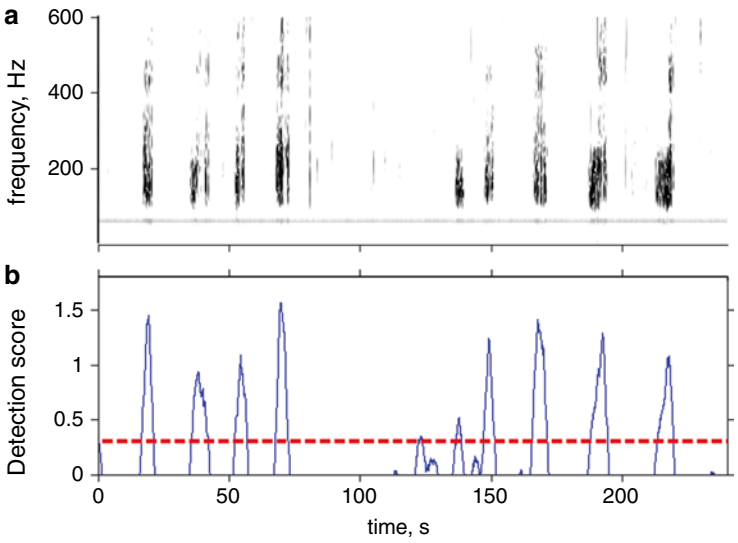

Fig. 15.3 Spectrogram and detection function for harbor seal "roar" vocalizations. Whenever the function exceeds the dashed-line threshold, a detection is registered

example of a *detection function* for harbor seal "roar" vocalizations computed from an audio signal. The decision criterion—a threshold in this case—is used to decide when to label the signal as having harbor seal sounds, or equivalently to trigger a *detection*. In this example, the decision criterion is used merely to decide the presence or absence of the desired sound. But in other contexts, decision functions may select from among several possible results, such as several types of calls. In such multi-way decisions, the decision criteria are correspondingly more complex, involving perhaps bounded $N-1$ dimensional hyperplanes in the N-dimensional space of features (Fig. 15.7b below).

Tradeoff in choice of threshold. Use of a threshold, or any two-way decision criterion, requires a numerical choice of that threshold, a choice that affects detector performance and involves a tradeoff. The tradeoff is between *wrong detections*, also known as *false positives, false alarms*, or *Type I errors*, and *missed calls*, also known as *false negatives* or *Type II errors* (Table 15.1). A higher threshold will result in fewer detections, reducing the probability of wrong detections but also raising the probability of missed calls, and vice versa for a lower threshold. The choice of threshold depends on the goal of the automatic detection. Some situations require detecting every call; this may be necessary when searching for an endangered species such as a right whale, or in real-time monitoring to ensure that no marine mammals are present in an area before doing something potentially harmful (e.g., blasting for construction of a harbor). In this case, a relatively low threshold is needed, with further checking of detections, either manually or with a classifier, to weed out the wrong detections. Other situations, like estimating population or population density, may require detecting only those calls that are relatively loud, but doing so as reliably as possible, with few false positives. For these situations, a relatively high threshold is needed. The section on performance measurement below discusses the setting of thresholds, including quantitative assessment of thresholds.

Table 15.1 Terminology use in describing detections

	Detected	Not detected
Desired vocalization	a	b
Anything else	c	d

The left side of the table indicates the truth about a set of calls—whether or not a given sound really is the desired call—while the top of the table indicates whether a given method detects the sound. The letters a, b, c, and d indicate the number of sounds of each type that occur
a = correct detection, true positive
b = false negative, Type error, II error, or miss
c = false positive, Type I error, or wrong detection
d = correct non-detection, true negative

$$\text{False positive rate} = \&\$\$\$;$$

$$\text{False negative rate} = \frac{b}{a+b}$$

$$\text{Precision} = \frac{a}{a+c}$$

$$\text{Recall} = \frac{a}{a+b}$$

Note that it almost never makes sense to speak of detecting or classifying *all* calls of the target species. Other than for captive-animal recordings, animals may be at widely varying distances from the hydrophone(s), with varying levels of interfering noise. A nearby loud call may be clear, but a sufficiently distant one is faint relative to background noise—i.e., it has a relatively low signal-to-noise ratio (SNR). In recordings made in the wild, there are always low-SNR calls at the limit of detectability and identifiability. This is true regardless of the detection and classification method used, including manual scanning.

Degree of automation. A closely related issue is the degree of automation needed. A fully automatic system is easy to use but probably unreliable. That is, it may require no supervision, but then no one notices if the detector/classifier makes wrong detections or misses calls—occurrences that are particularly likely if interfering noise in the background changes. At the other extreme is manual scanning—that is, a person manually checks all recordings by examining spectrograms or listening to the calls. This process is quite labor-intensive but is nevertheless useful when high confidence is required, as in the case of clearing an area of marine mammals before some potentially harmful action.

Most applications of automatic detection/classification fall somewhere between these two extremes. One popular technique is to check some subset of the detection/classification results to find the fraction that are wrong, then use this fraction to estimate the number of wrong detections in the full data set. In doing this, one must take care to examine separately those time periods when the fraction of wrong detections is likely to be different. This can happen either when the expected number of calls varies—which can happen because of migration or other movement, seasonal or diurnal changes in calling behavior, etc.—or when the background noise varies and thus alters the likelihood of a wrong detection—which can happen due to the appearance of interfering species' calls, changes in physical noise due to wind, waves, ice, etc., or changes in anthropogenic noise, like an increase or decrease in vessel noise. Another popular analysis method is to use automatic detection to find potential calls, then check *all* detections to determine which are correct. This can be useful when searching for a rare or endangered species, and can be combined with sampling of some time periods when no calls are detected to determine whether missed calls are an issue.

Desired level of specificity. How narrow a category of sounds must be detected? Different applications of automatic detection/classification will have different answers to this question, and will require different detection methods. At the most general level, one may wish to detect all possible marine organisms, as for a study that examines possible ecological and trophic interactions. At less general levels, to comply with the marine-mammal protection laws, one may wish to detect all marine mammal sounds present. One may wish to detect a certain taxonomic group—for example, detect all members of the family Ziphiidae (beaked whales). One may wish to classify sounds of a certain group defined acoustically, such as mid-frequency whistlers including dolphins, pilot whales, *Berardius* beaked whales, etc. One may wish to detect threatened and endangered species, either to study them or

to avoid possible harm to them. One may wish to detect a certain species, a certain call type, or at the most extreme level of specificity, calls of a certain individual.

These different levels of specificity require different approaches to detection and classification. For instance, finding sounds of all marine organisms requires a very general detection and classification system, such as a simple transient detector plus perhaps a classifier to remove known interfering sounds. A very specific task, like finding whistles of a certain single species of dolphin, may require a very specialized system: detecting all whistles in a certain frequency range, then measuring features of the whistles and using a classifier on the feature set to distinguish species. Finding clicks of odontocetes, a task at an intermediate level of specificity, can be done using detection of sounds that occur across a wide band of frequencies, plus further tests on the duration of these wide-band sounds.

Difficulty of detection or classification. The difficulty of a given detection task depends on several factors. One is the *call stereotypy*—that is, the degree to which different calls from the same or different individuals resemble one another. Highly stereotyped calls like blue whale "B calls" are relatively simple to detect and classify, and are amenable to the template-matching methods discussed below, while the highly variable units of humpback or bowhead whale songs are comparatively difficult (Fig. 15.4). Some call types may include both stereotyped and variable components, in which case it may be feasible to detect and classify only the stereotyped portion.

The type of call can also affect the choice of detection method and the difficulty of detecting and classifying it. For instance, click sounds of echolocating cetaceans require different detection techniques than the whistles of dolphins: Clicks can be detected using the simple time-domain methods discussed below, while whistles usually require a more complex frequency contour tracking method.

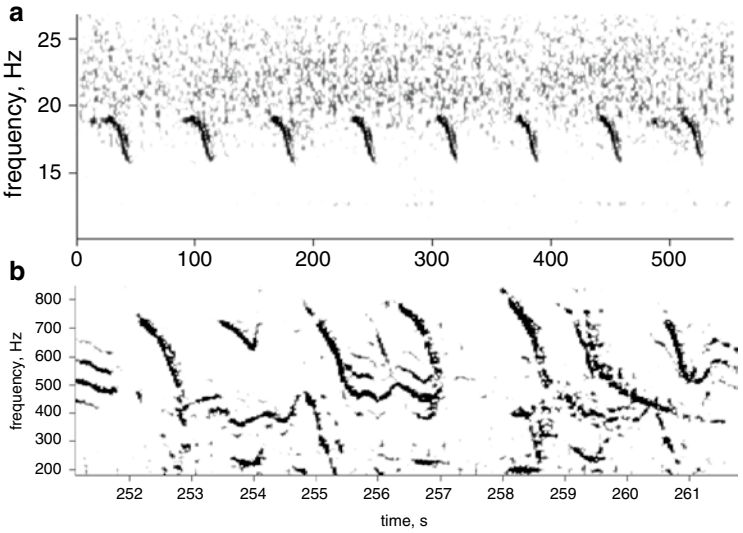

Fig. 15.4 (**a**) Highly stereotyped vocalizations of Atlantic blue whales. (**b**) Highly variable song units of bowhead whales

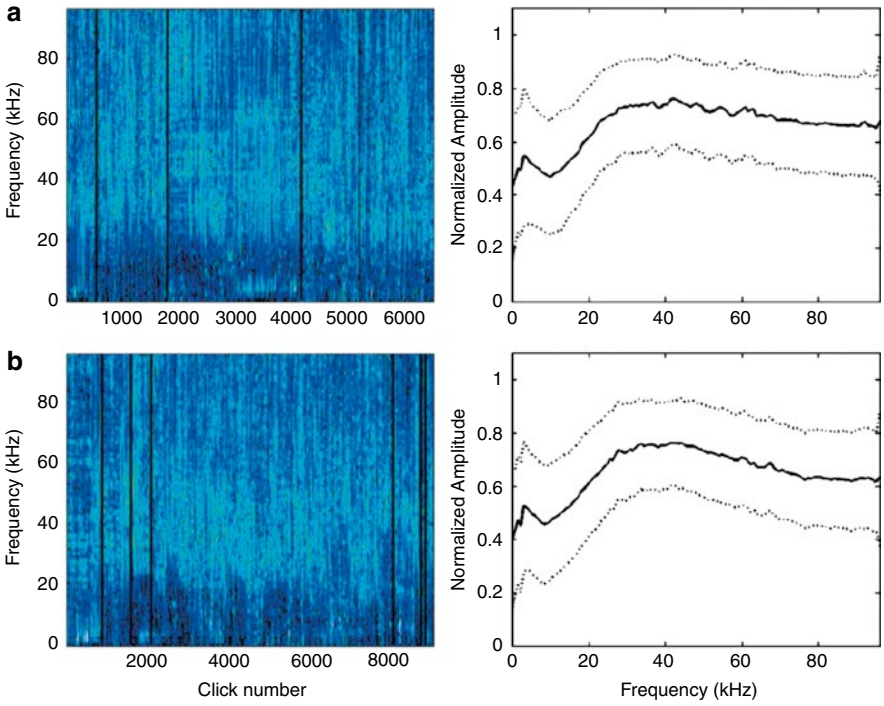

Fig. 15.5 Spectra of the clicks short-beaked common dolphins and common bottlenose dolphins. Spectra of individual clicks are plotted on the *left* with lighter tone denoting higher energy. (Note these are not continuous spectrograms, but rather compiled spectra of just clicks.) Averaged spectra are shown on the *right* by *solid lines* with *dotted lines* showing ±1 standard deviation. Their similarity makes separation of these two species difficult. Reprinted with permission from Soldevilla, M.S., E.E. Henderson, G.S. Campbell, S.M. Wiggins, J.A. Hildebrand, and M.A. Roch. 2008. Classification of Risso's and Pacific white-sided dolphins using spectral properties of echolocation clicks. J. Acoust. Soc. Am. 124: 609–624. Copyright 2008, Acoustical Society of America

Interfering sounds have a large effect on the difficulty of detection and classification. Masking by wide-spectrum background noise reduces the SNR of all calls, making detection and classification more difficult. Interference from other sounds in the environment can be even more of a problem, particularly when it has characteristics similar to the calls of interest. Most often similar sounds come from other species; cases in point are the vocal similarity of right and humpback whales (Mellinger et al. 2004), and the similarity between the clicks of common bottlenose dolphins (*Tursiops truncatus*) and short-beaked common dolphins (*Delphinus delphis*; Soldevilla et al. 2008), as shown in Fig. 15.5. Interference can also be nonbiological in origin. Indeed, in polar regions, the sound of ice cracking and rubbing has extreme variety, and is capable, for short time periods, of mimicking sounds of many different marine organisms. The general message here is *know your noise*. Noise, and its variation over time and space, have a large effect on the performance of detection and classification systems.

15.3.1 Conditioning

Signal conditioning refers to pre-processing a signal, or some representation of a signal such as a spectrogram, to prepare it for detection and classification. Some types of conditioning known as *normalization* are done to make an input signal more uniform, so that later stages of analysis have the behavior one might expect. For instance, a simple form of signal conditioning is to use automatic gain control to make an audio signal have a desired average sound level. Typically this involves calculating the moving average level (using some averaging time constant t_a), then dividing the signal by this average and perhaps multiplying by a constant to achieve the desired average level. The time constant t_a used in averaging should be chosen bearing in mind the call type to be detected or classified; using too small a t_a can make the averaging process silence the desired calls, while too large a one can make it fail to reduce background noise quickly, perhaps leading to poor performance at detection and classification. One rule of thumb is to use a time constant such that a new loud sound is reduced to half its original level in a period 3–5 times the duration of the desired call type.

Signal conditioning is also performed on spectrograms and other time–frequency representations. Often this is done for removing noise, or *de-noising*. One way to do this uses the same long-term averaging described above, but operates in each frequency bin independently. This technique, known as *pre-whitening* or *spectrum flattening*, has the benefit of removing long-duration, constant-frequency sounds such as vessel propeller noise and motor sounds (Mellinger et al. 2004). Another technique uses a wavelet transform to effect the de-noising (Kovesi 1999; Gur and Niezrecki 2007). Other forms of conditioning that are applied to spectrograms include image processing filters for various purposes. One type smooths edges in the image, so that frequency contours are easier to detect; it has been employed to detect right whale calls (Gillespie 2004). Other examples of image processing filters include the opening and closing operators which join areas that are almost connected and smooth away rough edges respectively. These have been used in the recognition of both baleen (Mathias et al. 2008) and odontocete (Mallawaarachchi et al. 2008) tonal calls.

15.4 Detection Methods

The most widely used detection methods are reviewed here; a similar review of classification methods follows. We use *input signal* to mean the sound signal in which we wish to find calls of interest, *detection function* to mean a function of time that reflects our belief that the desired sound is present at any given time, and *threshold* to mean the level above which the detection function must rise to indicate a detection. The most straightforward detection methods operate in the time domain, i.e.,

using the time series signal itself rather than another representation like a spectrum or spectrogram, and we review them here first.

Matched filtering is a template-matching method in the time domain. It consists essentially of cross-correlation of a fixed template, the *kernel*, with the input signal. The kernel is normally a copy of the call of interest, either a very clear recording or a synthesized version of the signal. The reason for needing a very clear version is that any noise in the kernel adds to noise in the detection function, increasing its error rate. Matched filtering has a long history in detection theory, having been used to detect radar reflections during World War II; it is the optimum linear filter for detecting a known sound in the presence of white Gaussian noise. "Known" in this context means that the waveform (time series) of the target sound is known exactly. Although animal calls are never "known" in this sense, as there is always some variation from one animal sound to the next, matched filtering is still useful for detecting highly stereotyped calls. It has, for instance, been used for detecting the B calls of blue whales (Stafford et al. 1998) and for discriminating the clicks of individual sperm whales (Gillespie and Leaper 1996). Matched filtering works less well when there is variation between calls, or when the background noise is not white—as when the sound contains significant vessel noise. Urazghildiiev and Clark (2006) present a method for detecting right whale calls by matching many possible templates in parallel.

Band-limited energy summation consists of simply using the level of the input signal within a fixed frequency band as the detection function. The waveform is bandpass-filtered to leave only the desired portion of the spectrum, so that sounds in this band result in increases in amplitude of the detection function (the filtered signal), and a threshold is applied to the result. This method is fairly general, in that it detects *any* sounds within the desired frequency band (though further processing of the detection function can be performed, as explained below, to restrict which suprathreshold events are considered detections). It has been used most often for detecting echolocation clicks of odontocetes, as for example for detection of sperm whale clicks (Gillespie 1997; Mellinger et al. 2004). Variants of this method have been developed for discriminating the desired clicks from those of other species present. Energy ratios between a band of interest and a neighboring band where energy is not expected (Au et al. 1999) have been used. A method known as the Energy Ratio Mapping Algorithm (ERMA) optimizes the selected frequency bands to distinguish the target species' clicks from expected clicks of other species in a survey area. The two corresponding bandpass filters are both applied to the input signal in parallel, and the ratio of these filters' output in combination with a Teager–Kaiser energy operator is used as the detection function (Klinck and Mellinger 2011).

A large class of detection methods is based on time–frequency representations of the input signal such as the spectrogram. Other time–frequency representations are sometimes used or suggested, including wavelets and the Wigner–Ville distribution, though spectrograms remain by far the most widely used in bioacoustics. Qian and Chen (1999) provide an overview of these other representations. Wavelets have been used two ways: Directly, in that the wavelet coefficients provide the input feature vector to a classification system, and indirectly, in that the features are derived from

the wavelet coefficients. Conversion of a signal to a time–frequency representation can make it simpler to detect sounds with particular time–frequency characteristics, including manual detection, and also makes it simple to apply conditioning techniques to equalize or "whiten" the long-term spectrum of the signal (Mellinger et al. 2004). This has the effect of reducing the effect of long-duration noise sources such as vessel sounds, wind and wave noise.

The Hilbert–Huang transform has been used to detect and analyze cetacean sounds. This transform, similar in spirit to a wavelet analysis, consists of decomposition of the signal into a "mode function," which is calculated from envelopes of the successive maxima and minima of the waveform, and the residual that is left after subtracting the mode function from the original signal. The decomposition is then repeated on the residual using a different, orthogonal mode function, and the whole process is iterated until the residual becomes sufficiently small. The result is a set of mode functions that describe the original signal. Adam has had success using this technique to analyze sperm whale clicks (Adam 2006a, b) and to track killer whale whistles (Adam 2006b, 2008).

The simplest of time–frequency methods is similar to band-limited energy detection: The detection function is simply the sum of spectrum values in a given frequency band—i.e., in the appropriate bins of the spectrum. This method has the same advantages and disadvantages as the similar method in the time domain discussed above, except that noise removal via spectrogram conditioning is possible.

Many animal sounds are composed of frequency contours—narrowband tonal sounds that change frequency over time. Such sounds include whistles of many odontocetes, moans of mysticetes, and trills of some phocid seals. Such sounds are typically detected using methods that find a peak in the spectrogram frame (spectrum) at the start of the contour, then track that peak over time in successive frames. If the peak is sufficiently high above background noise and persists for a sufficient duration, a detection is registered. Methods employing these ideas have been used to analyze whistles of bottlenose dolphins (Buck and Tyack 1993) as well as moans of baleen whales (Mellinger et al. 2011). The advantage of these methods is that they detect frequency contours of all shapes and sizes within a specified frequency band; this is also their disadvantage, because if there are interfering frequency contours in this band, the methods typically detect them as well.

A number of other tonal detection methods are based on processing spectrogram energy. Gillespie (2004) presented another method for detection of frequency contours based on edge-detection techniques from the field of image processing. The spectrogram is smoothed to eliminate speckle, and the outlines of sounds are found using an edge-detection algorithm. If the contour is longer than a specified minimum duration, it is then subjected to further analysis to determine whether it is from the desired target species. Several groups have used Bayesian filtering, where the spectral peaks observed during the detection process are used to update a posterior distribution of where the next peak in a tonal might occur, and a statistic of the distribution (e.g., the mean) is used as a point estimate. This was first reported by Mallawaarachchi et al. (2008) and White and Hadley (2008) with Kalman and particle filters respectively. Roch et al. (2011b) showed that more advanced particle

filters could perform well in complex auditory scenes with many animals producing calls simultaneously. Their work also considered delaying decisions about crossing whistles until groups of intersecting whistles were entirely detected, permitting information from both sides of the crossing to be used. Finally, Kirshenbaum and Roch (2013) applied image-processing based ridge detection algorithms.

Spectrogram correlation is a template-matching method in the time–frequency domain. As with matched filtering, it involves cross-correlation of a kernel with the input signal, only this time the input signal is represented as a spectrogram. The kernel can either be synthesized or generated from a recording; synthetic kernel generation methods generally include mechanisms to detect the calls of interest while rejecting interfering calls that occur simultaneously. Spectrogram correlation has the advantages and disadvantages of template-based methods: the method permits high specificity with respect to call type, but detection performance declines when calls vary too much from the template. The method generally allows for more variation in calls than matched filtering does, and Mellinger and Clark (2000) present a method for handling variation in the timing of successive parts of calls. Spectrogram correlation has been used principally for detecting stereotyped calls of baleen whales, including blue whales (Mellinger and Clark 1997), right whales (Munger et al. 2005; Urazghildiiev et al. 2009), and sei whales (Baumgartner et al. 2008). While many baleen whale calls are highly stereotyped, some call characteristics have been shown to experience drift over time. An example of this is blue whale B calls in the Northeast Pacific, whose dominant frequency has been shown to decline by nearly a third over a period of over 40 years (McDonald et al. 2009). This has led some researchers (e.g., Oleson et al. 2007) to develop kernels specific to certain time periods.

Another spectral method uses phase information to detect echolocation clicks. Kandia and Stylianou (2006, 2008) show that the position of a delayed unit impulse can be predicted by the group delay (negative derivative of the signal's phase spectrum), and the average over frequency for the group delay function similarly predicts the delay of rapidly decaying functions such as an echolocation click. They propose a method to estimate the slope of the group delay and use sets of sliding windows to detect when an echolocation click is at the origin of a window. Negative phase slopes are indicative of an impulsive sound far from the start of the frame. As the window slides, a negative-to-positive zero crossing of slope indicates that an echolocation click is at the origin. This method is robust to high levels of background noise and is relatively nonspecific, detecting all short-duration impulsive sounds such as odontocete echolocation clicks.

Finally, the detection can be based on the entropy estimated from spectrogram frames. These methods estimate a statistic called the Shannon information entropy that measures the amount of information in the signal. Portions of an input signal having marine mammal calls contain more information, and so the entropy statistic over time can be used as a detection function. This method is very general, detecting a wide variety of cetacean and pinniped sounds (Erbe and King 2008). This generality is both its strength and its weakness; it would be most useful for detecting the

presence of any marine mammal, but not useful for detecting a certain species or call type. Entropy methods have also been use to analyze the information content of humpback whale songs (Suzuki et al. 2006; Miksis-Olds et al. 2008).

While most of the detectors described so far operate in the time–frequency domain, detectors for both tonal and impulsive calls can operate on time-domain signals. For a tonal signal $x[n]$, its instantaneous frequency can be estimated from an analytic signal $y[n]=x[n]+jH(x[n])$ where H denotes the Hilbert transform and $j=\sqrt{-1}$. The instantaneous frequency is defined as

$$f_i(t)=\frac{1}{2\pi}\frac{\mathrm{d}\phi(t)}{\mathrm{d}t}$$

where $\phi(t)$ is the phase of the analytic signal $y[n]$. This can be interpreted as the mean of the changing spectrum at time t (see Boashash 1992 for a thorough discussion of instantaneous frequency), and the goal of the time domain detectors discussed here is to track how instantaneous frequency evolves over time. Ioana et al. (2010) modeled the instantaneous frequency by analyzing short segments in which the instantaneous frequency could be modeled by a series of piecewise polynomials. An alternate process proposed by Johansson and White (2011) tracked tonal calls by optimizing a set of notch filters over time. The filter parameters follow the instantaneous frequency and permit recovery of the whistle. The developments in this area are interesting and merit further investigation; however at the time of this writing, there remain significant challenges in dealing with complex and noisy data sets.

The Teager energy operator (Kaiser 1990) is a short-time energy estimation method used in the bioacoustics community for detecting brief calls such as echolocation clicks. Proposed by Teager and developed by Kaiser, it is sometimes referred to as the Teager–Kaiser energy operator and estimates energy based on three samples. The energy is based on the energy required to *generate* simple harmonic motion in a mass-spring model. The operator estimates the energy needed to excite such a system, which is proportional to the square of the amplitude and frequency of the measured signal. Kaiser showed that for a variety of non-harmonic human speech signals, the Teager energy operator still gave very good indications of where energy was present. Kandia and Stylianou (2006) were the first to propose using the Teager energy operator to detect echolocation clicks of sperm whales. Due to the broadband nature of odontocete echolocation clicks whose peak frequencies are typically in quieter portions of the spectrum, the high frequencies tend to result in strong rises of Teager energy (Fig. 15.6). Kandia and Stylianou showed that the skewness of the Teager energy distribution could be used to efficiently determine whether or not an echolocation click existed over a given window. When clicks were present, an energy growing algorithm permitted the recovery of clicks. Echolocation clicks violate the assumptions of the model (non-harmonic signal, and the estimation error increases greatly when the frequency is greater than $1/8^{\text{th}}$ of the sampling rate), yet Kandia and Stylianou showed empirically that the Teager energy was effective for detecting echolocation clicks.

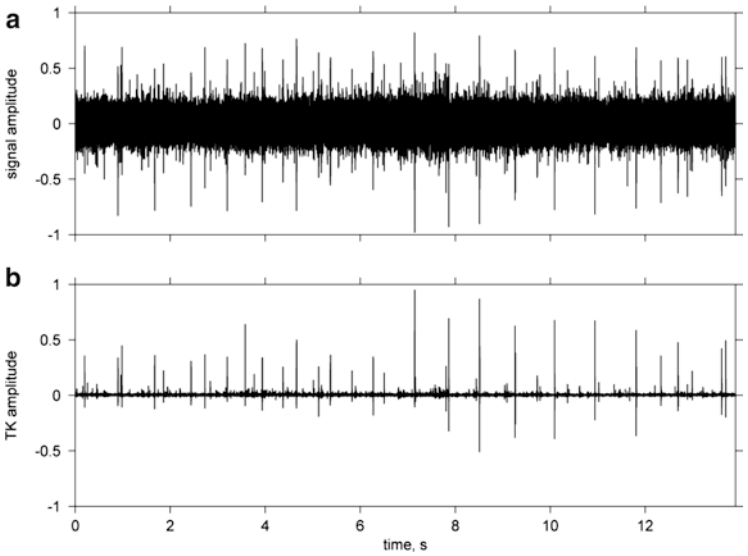

Fig. 15.6 (**a**) An acoustic signal containing sperm whale clicks. (**b**) The result of applying the Teager–Kaiser operator to this signal. Note that the clicks stand out much more above background levels here than in the acoustic signal

15.4.1 Detection Function Processing

The methods mentioned above produce a detection function, which must then be analyzed to find discrete *detection events*—times when detections, and hopefully calls, occur. The simplest way to do this is simply to register a detection event whenever the detection function surpasses the threshold, but a number of refinements to this method are often helpful.

Multipath rejection. Marine bioacoustic sounds often reach a hydrophone by multiple paths—echoes off the sea surface or floor, multiple refractive paths within the water column, or some combination of these. Usually one desires to ignore these multiple arrivals and register only one detection event per call produced by the animal. A simple means to do this is to have a short *refractory period* after a detection event, such that no further detections are possible within this period. The length of this period depends on the geometry of the multiple paths between source (the animal) and receiver (hydrophone). This rejection method is effective, but it runs the risk of rejecting other calls, perhaps from nearby conspecifics, that happen to arrive during the refractory period. To avoid this, one can reject other calls within the refractory period if the absolute value of the normalized cross-correlation of the first arrival and a later arrival is above certain amount; this is usually effective because multipath arrivals of a call are typically (though not always!) highly similar in structure. The absolute value operation is needed because of the sign change (phase inversion) that happens to acoustic pressure waves when they reflect off the water's surface.

Jitter rejection. The detection function typically contains a significant amount of jitter—variation on a very short time scale. This jitter can cause the detection function to cross the detection threshold several times while rising above or falling below that threshold in the long run, possibly triggering multiple detections. Two approaches to handling this are effective. One is to *smooth* the detection function—to take an average, or perhaps a weighted average, of every group of n samples. Here n is essentially a time constant that determines the time scale over which smoothing occurs. Heuristically, it has been found to work well to use a time constant roughly equal to or less than the duration of calls to be detected, depending on the detection method used. Smoothing lowers the height of detection function peaks, which presumably occur when a call is present, so it is necessary to adjust the detection threshold when using it; fortunately, it also tends to reduce the height of the detection function when calls are *not* present as well, so non-calls are still rejected. The other method for handling jitter is to register a detection event only for a local peak in the detection function—i.e., when the detection function is larger than all other values within a neighborhood of a certain duration. As above, the duration should be approximately the duration of the call to be detected.

Enhancing energy localization. The Teager energy operator has been used by several groups for detecting echolocation clicks of odontocetes (e.g., Roch et al. 2008) and with varying modifications, such as signal preconditioning with high-pass filtering (e.g., Bénard and Glotin 2010; Gervaise et al. 2010; Soldevilla et al. 2008). The technique has also been applied to the output of detection algorithms (Klinck and Mellinger 2011) to find regions of high short-time energy.

Adaptive threshold. The threshold need not be constant. It can be beneficial to calculate a long-term average of the detection function and adjust the threshold height to it. This is especially helpful in two cases for which the variance of the detection function changes over time. First, the performance of time-domain methods can suffer because of a change in background noise; this essentially raises the height of peaks in the detection function, including unwanted peaks due to noise or interfering sounds. Second, even spectrogram-based methods that pre-whiten the background noise can have increased variance in the detection function as a result of heightened noise, and these changes in variance can again trigger false detections. Having the detection threshold change in response to changes in the variance of the detection function (Gillespie 1997) helps solve both of these problems.

Detecting regular calls. Bioacoustic sounds that occur at regular intervals can be detected by methods that are sensitive to regularly occurring peaks in the detection function. One way to do this is by taking successive frames of the detection function—successive fixed-size sequences of samples of it—and computing the autocorrelation of each frame. Peaks in the autocorrelation between the times (lags) corresponding to known call intervals then indicate regularly occurring calls. This method has been effective at detecting regular sounds that are too faint to detect directly in the spectrogram. Many cetaceans use regularly occurring vocalizations at some point in their life cycle; this method has been used on songs of fin whales (Mellinger et al. 1994), pulse trains from minke whales (Mellinger and Clark 1997), and clicks from sperm whales (Mellinger et al. 2004).

15.4.2 Classification

After deciding what to classify and selecting an appropriate feature set, one must decide what method will be used for classification. One of the most important lessons for those wishing to classify data is that there is no one best method for classification. In fact, the aptly named "No free lunch" theorem (Duda et al. 2001) shows that this is the case. Consequently, it can sometimes be useful to try multiple classification techniques on the same data set. That said, no classifier will help when there is a poor feature set, and selecting good features is one of the most critical steps in developing an effective system. Formally, the task of a classifier is to assign a label to a set of features derived from phenomena that one wishes to classify.

Classification systems can be broadly divided into generative and discriminative techniques. *Generative classifiers* learn how features associated with each class are distributed and decide the class label for a new instance (animal call) based on some measurement of similarity to the training distribution. In contrast, the designers of *discriminative classifiers* do not concern themselves with how features are distributed, but rather how to separate classes. Figure 15.7 shows a sample of features derived from echolocation clicks of Risso's dolphins and Pacific white-sided dolphins. These two species are readily distinguishable acoustically (Soldevilla et al. 2008) and one can see a very good separation in even the first two cepstral feature vectors here. The left plot shows an example of a simple Gaussian classifier, where the shapes of the two multivariate Gaussian distributions have been estimated to maximize their fit to each species' training data. Likelihood contours are plotted about the means of the two distributions. To use such a classifier for a call, one calculates the feature vector(s) for the call and determines which distribution would have the highest likelihood for that vector(s). In contrast, the right plot shows a line perpendicular to the separating hyperplane resulting from linear discriminant analysis. Test vectors are also projected onto the line, and classified based upon where they lie on that line. The boundary is roughly the midpoint between the means of the projected training vectors.

Classifiers can be thought of as producing a static partitioning of the feature space. Figure 15.8 shows the partitioning for a subset of a two-dimensional feature space in a three-class species identification problem. This example was produced with two-dimensional click features and a generative classifier.

Discriminative methods have the advantage that they attempt to optimize the classification decision, and many have argued that these techniques are in general more appropriate for classification. A caveat to this is that the training data must adequately characterize the separation boundary. As an example, if one were to build a "detector" for Risso's dolphins using only the toy data sets of Fig. 15.7 (not recommended), *all* other species would have to fall on the correct side of the boundary. In contrast, a generative model could set a threshold such that anything sufficiently distant from the training distribution would be rejected.

In addition to considering classifiers as generative or discriminative, one must also consider whether or not the goal of the classifier to is to learn known categories

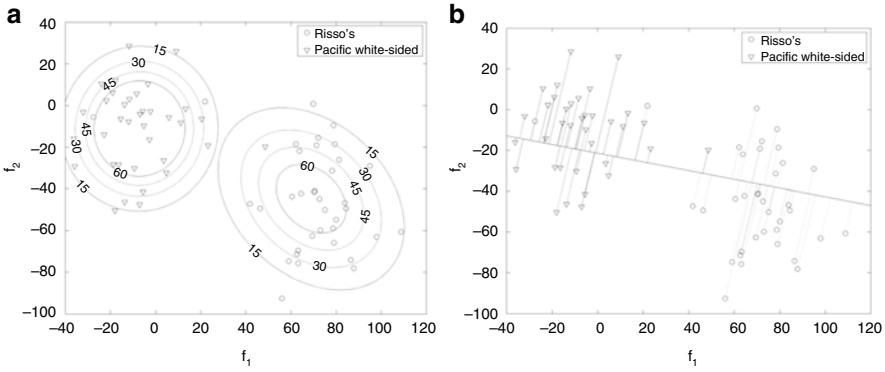

Fig. 15.7 Comparison of two classifiers trained on echolocation click features from Risso's and Pacific white-sided dolphins. (**a**) The generative classifier on the *left* models the click features as a Gaussian distribution for each species. *Contour lines* show likelihood values scaled by 10^5 for readability. (**b**) The discriminative classifier on the *right* shows projection onto a *line* selected by linear discriminant analysis

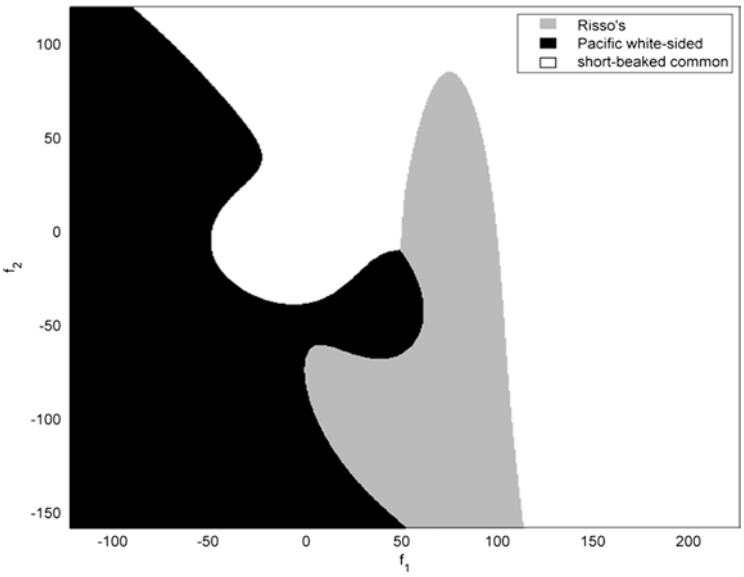

Fig. 15.8 A three-class species identification problem for Risso's, Pacific white-sided, and short-beaked common dolphins showing that classifiers induce a partitioning of the feature space. As in the previous figure, a classifier was trained from two-dimensional feature data derived from echolocation clicks. This example introduces a third species and uses data from several sightings. A two-mixture Gaussian mixture model (described later in this chapter) was trained for each species. Rather than plotting training vectors as in Fig. 15.7, this plot shows the species that would be selected for any test vector within the range of the plot. Decisions are made by selecting the species associated with the model with the highest likelihood

or to discover categories on its own. When the class labels are provided in the training process, the classifier is called a *supervised* learner. *Unsupervised* learners determine groups based solely on properties of the data, and it is up to the human analyst to determine if the groups carry any significance.

Both the Gaussian classifier and linear discriminant analysis are examples of models used for classification. In both cases, the training data is used to determine parameters for an algorithm that distinguishes between different types of feature vectors. In the Gaussian classifier, maximum likelihood estimation could be used to show that the Gaussian distribution which maximizes the probability of each class's training data is the sample mean and covariance. In the case of linear discriminant analysis, the hyperplane is chosen so as to maximize the separation between points of the different classes when they are projected onto a separating line. Fitting a classifier depends upon the type of classifier, but generally it involves maximizing (or equivalently minimizing) some statistic of the training data. After fitting, the model's performance is evaluated (see details later in this chapter). In most cases, the eventual goal is to have enough confidence in the classifier's decisions to apply it to field data where the result is not known. Except in the case of simple problems, no classifier will have perfect performance, and one needs to understand the classifier's performance to use it effectively in a bioacoustic study.

15.4.3 What Is the Right Type of Classifier?

Selection of an appropriate classifier for a call depends upon numerous issues. The analyst must consider the characteristics of the calls to be classified (e.g., is it a long frequency-modulated call such as a moan or whistle that varies over time or a short echolocation click?), whether the goal is classification or understanding what features are important for classification. Finally, the analyst must consider how much expertise they or others have working with available software packages or developing them on their own.

From a theoretical perspective, classification errors are composed of several different components. The Bayes error (also called Bayes rate) is the classification error that would occur with an optimum classifier for a given feature space and distribution of features. Unfortunately, real-world classifiers do not typically achieve the Bayes rate, which assumes that one knows the exact distributions of the classes being modeled and that features are measured without error. There are many factors that can corrupt feature vectors, including ambient noise, propagation effects such as dispersion and echoes, measurement error, and a host of other factors that serve to distort the feature vectors associated with the call being measured. Error above the Bayes rate is composed of two components, bias and variance. The *bias* is a result of structure imposed by the type of classifier used. Manning et al. (2008) give the example of classifying data that is separated by a nonlinear boundary. Using a family of classifiers capable only of linear separation would be likely to produce a high bias, as they would not be able to construct the appropriate nonlinear boundaries between classes. In contrast, *variance* is related to how sensitive the classifier is

to variation in a training set. A classifier that produces very different results when given slightly different training data exhibits high variance.

The number of parameters in a model, or its order, is related to bias and variance (Hastie et al. 2001). When the model order is low, the bias tends to be high. As the order is increased, bias decreases and the error rate on the training set (but not necessarily on an independent test set) will decrease. Unfortunately, as one achieves a better and better fit of the training data set, one learns the idiosyncrasies of that particular data set rather than characteristics of the population from which the sample was drawn. This *overfitting*, or *overtraining*, of the training data results in a high variance and a poor error rate when given different data to classify. This is known as the bias-variance tradeoff and in general the search for an appropriate classifier is an attempt to find the model that optimizes the balance between the two types of controllable error.

Many classifiers are designed to discriminate between only two classes. While this may appear to be limiting, it does not pose serious challenges. To solve multiclass problems with two-class classifiers, one typically trains one classifier per class, with each one learning one of the categories (e.g., blue whale D call) versus all other categories. To classify a new call, it is evaluated by each classifier, and the one with the best response is selected.

In the next several sections, several types of classifiers are discussed. They can broadly be divided into parametric and nonparametric classifiers. While all classifiers have parameters, such as thresholds, *parametric* classifiers attempt to fit parameterized statistical distributions such as Gaussian distributions. A *nonparametric* classifier, in contrast, has no assumptions about an underlying distribution for the data. The tour concludes with a brief overview of unsupervised learning. Throughout this discussion, the goal is to provide the reader with an intuitive feel as to how each classifier functions as opposed to the complete understanding that one would require to implement the method. The discussion is far from exhaustive and should not in any way be considered a complete account of machine learning techniques. There are several excellent books on machine learning and the interested reader is referred to Duda et al. (2001), Hastie et al. (2001, 2009), and Mitchell (1997).

15.4.4 Nonparametric Classifiers

For highly stereotyped calls, there are a number of simple but effective methods that are based on template matching. The central concept for template-matching classifiers is that the call is not expected to vary significantly from the examples, or templates, to which they are to be matched. The previously discussed matched filters and spectrogram correlation methods can both be seen as examples of nonparametric classifiers. A limitation of both of these methods is the inability to account for changes in time variability in a signal. A method of permitting nonlinear variation in the timing of call production is the use of dynamic time warping, a technique used in early speech recognition systems (Rabiner and Juang 1993). In dynamic time warping, one aligns feature vectors from a template call to those of a test call. The method uses a dynamic programming algorithm to efficiently find optimal

pairings between the feature vectors of the template and test call. This permits nonlinear alignment, or speeding up and/or slowing down portions of the call. Dynamic time warping has been used for recognizing signature whistles of bottlenose dolphins (Buck and Tyack 1993) adapted to model timing between piece-wise spectrogram correlation of components of bowhead whale song (Mellinger and Clark 2000), killer whale calls (Brown and Miller 2007), and used as part of a system to cluster delphinid whistles (Deecke and Janik 2006; see discussion of unsupervised methods below).

A final type of template method is nearest neighbor search. This technique accounts for variability in a template by allowing many examples of templates, each stored with a class label. When an acoustic sample is presented to be classified, a similarity metric is used to determine which k templates best match the sample (Duda et al. 2001). It is up to the practitioner to choose an appropriate value of k. The class labels of these k "neighbors" are examined, and a majority vote is used to decide to which class the sample belongs. While such a technique would seem to be computationally expensive, considerable effort has gone into computational methods to perform this task in a reasonable time even when there are a large number of examples. The well known k-means algorithm (Mitchell 1997), also called vector quantization, can be thought of as an approximative variation of nearest neighbor search. Training data are clustered and clusters are labeled according to the most frequently occurring class in the cluster. Instead of searching for the k nearest neighbors, cluster means represent the data, and a search is made for the closest cluster mean, resulting in significantly reduced search time.

As mentioned above, linear discriminant analysis can be used to find separating hyperplanes, and many more sophisticated methods uses trees or networks of linear discriminant classifiers. While linear discriminant analysis cannot model complex partitions of the feature space, choosing the right features can make them quite effective. A particularly elegant example of this can be found in the work of Gillespie et al. (2013), where the authors split whistles into segments and extracted simple features from the segments (e.g., mean, slope, curvature) and generated distributions of these statistics based on samples from many segmented whistles. Statistics of these distributions were computed and used as feature vectors that were classified by linear discriminant analysis.

Decision tree classifiers use a series of questions about feature values, such as "Is the center frequency of an echolocation click within a certain range?" The first question forms the root of the tree, with subsequent questions fanning out like the branches of the tree. Much like the popular children's game of 20 questions where a player attempts to determine of whom or what their opponent is thinking using yes or no questions, these systems partition the feature space into rectangular regions, or hypercubes. Each hypercube is either labeled by a class or further subdivided by another question. Decision trees can be seen as a form of rule-based system, and when a human's knowledge and intuition is used to construct the rules we refer to this as an expert system. Madhusudhana et al. (2009) developed such a system for the classification of B and D calls produced by blue whales (*Balaenoptera musculus*). Unfortunately, the rules used by humans are not always easy to quantify nor can they be generalized easily when new classifiers are desired.

Alternative forms of decision trees determine which questions to ask automatically. The systems examine the possible rules that could be used to split the dataset at each point and then select the rule that best separates the data set. An impurity measure is used to evaluate the quality of each potential split. Several impurity metrics are commonly used, but the general idea is to determine if the proposed split results in improvements to the classification error or to an information theory metric such as cross-entropy (Hastie et al. 2003). This process is repeated recursively on each split until the nodes contain only a single class or some metric is met. Tree classifiers frequently overfit the data. Consequently, a critical step for most tree-classifiers is to prune some of the lower level splits after the tree has been trained. Perhaps the two best known tree classifiers are classification and regression trees (usually referred to by the acronym CART), and the C 4.5/C 5.0 algorithms (Hastie et al. 2001). CART has been applied to the task of determining which species of odontocete has produced a set of whistles by Oswald et al. (2007). Tree based classifiers offer the advantage over other types of classifiers that it is typically easier to understand how the algorithm made its decision.

There are a large number of classifiers that are covered under the name of "neural networks," or connectionist networks as they are sometimes called. One of the most popular of these in the bioacoustics community is the back-propagation neural network, which consists of interconnected nodes called perceptrons. Each node is capable of separating the data linearly, but when they are combined, the network is capable of performing nonlinear separations of data (Lippmann 1989). As shown in Fig. 15.9, the components of an input feature vector $\vec{f} = [f_0, f_1, \ldots, f_n]$ are fed to an input layer of perceptrons.

Each node in the input layer projects the feature vector onto a line (similar to linear discriminant analysis). The results of these classifications are distributed to

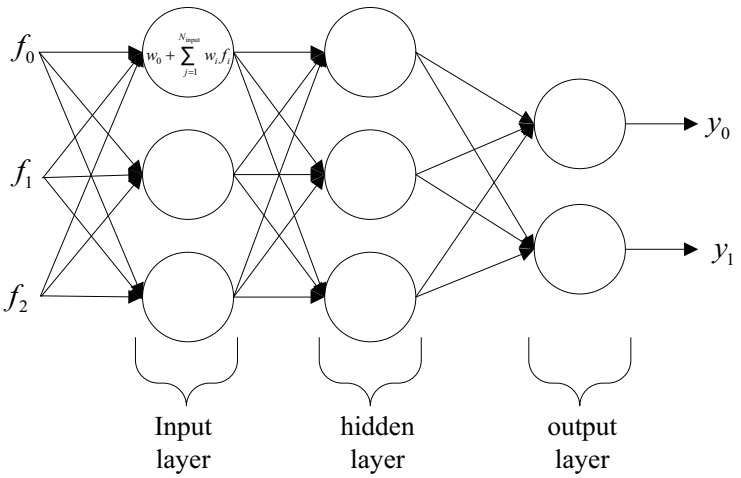

Fig. 15.9 A feedforward neural network with one hidden layer. Each component of the feature vector $\vec{f}$ to be classified is presented to the input nodes of the classifier. The results are propagated through the network and the output vector $\vec{y}$ contains the classification result

the nodes of a hidden layer where the process is repeated using the previous layer's output as input. In principle, multiple hidden layers are possible, but typically only one is used. With enough nodes and training data, a single hidden layer can model any input–output relation, though the number of nodes needed might be large and presents a risk of increasing the variance (overtraining). The hidden layer delivers values to the output layer whose outputs are used in the classification decision. In the earlier example of distinguishing echolocation clicks of Risso's dolphins from those of Pacific white-sided dolphins, one could train the network to output a value close to 1 on y_0 when the decision is that the click was produced by a Risso's dolphin and a value close to 1 on y_1 otherwise (Fig. 15.9).

Training is an iterative process, where the node parameters are adjusted at each iteration to make the output agree with the class of the training samples. A parameter called the learning rate controls how aggressively the node parameters are updated. When the learning rate is high, nodes are adjusted by large magnitudes, but large adjustments may skip over a good parameter set. Lower learning rates increase the number of iterations required but are less likely to "overshoot" a good set of node parameters. A common strategy is to start with a large learning rate and to decrease it over time. Due to the ready availability of software and generally good performance, neural networks have been used extensively for cetacean bioacoustics. Examples of this method used on various cetacean discrimination tasks include Deecke et al. (1999), Houser et al. (1999), and Potter et al. (1994).

A final form of nonparametric classifier is Vapnik's support vector machine (Burges 1998; Cristianini and Shawe-Taylor 2000). Support vector machines (SVMs) are linear classifiers which have the potential to separate nonlinear data by projecting them into a higher dimension where linear separation is possible. The separating hyperplane is chosen by minimizing an empirical risk function under a 0–1 loss rule. The result of this is that the hyperplane is selected so as to maximize the distance between points of different classes. To account for cases where the training data is not linearly separable in the higher dimension, a user settable penalty parameter is introduced that increases the value of the optimization function when points fall on the wrong side of the hyperplane. When using a support vector machine, one must also decide what kernel to use. Kernel functions provide weight, or support, for a local neighborhood about a point, and common choices for kernels (Hastie et al. 2001) include polynomial, radial (Gaussian) basis, and neural network (sigmoid) functions. Kernels typically have parameters, and the SVM's performance will thus be a function of the penalty, kernel function, and kernel parameters. Support vector machines have been used to distinguish odontocete species by their echolocation clicks (Jarvis et al. 2008; Roch et al. 2008).

15.4.5 Parametric Classifiers

Parametric classifiers attempt to model the *posterior distribution* of a class ω (e.g., species, group call type) given a feature vector x as evidence: $P(\omega|x)$. Decisions are made using the Bayes decision rule, which selects the class ω from the set of all possible classes Ω that has the highest posterior likelihood:

$$\arg\max_{\omega \in \Omega} P(\omega \mid x).$$

When the posterior distribution is accurate (this is rarely the case), this decision rule minimizes misclassification error. Direct estimation of $P(\omega|x)$ is difficult, but Bayes rule can be used to rewrite this probability as

$$P(\omega \mid x) = \frac{P(x \mid \omega) P(\omega)}{P(x)}.$$

$P(x|\omega)$ is referred to as the class-conditional likelihood and $P(\omega)$ is the prior probability. The *prior probability* is the probability that the next observation will come from class ω and is frequently unknown. In such cases, a *non-informative prior*, or uniform distribution, is used. The class ω is decided by using the class associated with the model that produces the highest posterior probability. As $P(x)$ is constant in the denominator above, it will not affect the maximum posterior probability and can be safely ignored, as can $P(\omega)$ when a non-informative prior is used.

It is possible to train parametric models to be discriminative classifiers. Doing so requires consideration of model parameters for different classes simultaneously. One example of this is maximum mutual information estimation, a technique that attempts to maximize the mutual information between training vectors and their associated class. When this is done, the object of training is to maximize the ratio of the correct class probability to that of a statistic of the competing models. A drawback of this technique is that parameter estimation becomes more difficult, and one typically must turn to methods such as gradient descent (Huang et al. 2001).

As a consequence of the difficulty of discriminative training, many parametric classifiers focus on maximizing the class conditional likelihood with respect to their training data. While many parametric classifiers exist, discussion will be limited to the two that are most prevalent in the bioacoustics literature: Gaussian mixture models and hidden Markov models.

Gaussian mixture models (GMMs) consist of a set of N Gaussian distributions scaled by a factor such that integration over the entire feature space still sums to one. These models are quite flexible and can model most distributions. Straightforward maximum likelihood techniques are not possible as one cannot attribute each training observation to a specific mixture. An application of the expectation–maximization algorithm (Moon 1996) permits a two-stage iterative process to create a model. In the first stage, the current model parameters are used to determine the expected associations between observations and mixtures. Using the expected values, a new maximum likelihood estimate is obtained. Convergence is guaranteed, and GMMs have been used for species identification for delphinids (Roch et al. 2007, 2011a), identification of killer whale calls (Brown and Smaragdis 2009; Shapiro et al. 2011) and in terrestrial bioacoustics for bats (Skowronski and Harris 2006).

With the exception of dynamic time warping, previously discussed classifiers are unable to exploit the temporal structure of the call. Hidden Markov models (HMMs, Rabiner 1989) provide a method to recognize calls that have similar structure but differ in the timing of the components. The fundamental concept that lets HMMs represent temporal evolution is that of a state. Each model consists of

several states together with probability distributions for transitioning from one state to another. Each state models the distribution of features (frequently using a Gaussian mixture model) that occur in that state. The model learns both the state distributions and the likelihood of transitioning between states. Like the aforementioned Gaussian mixture model, information needed to compute a maximum likelihood estimator during training is not available, and the expectation–maximization algorithm is used. Both training and testing require the examination of many possible paths through the model, and dynamic programming algorithms permit this to happen in a tractable manner. These models have been used to determine group association by analyzing delphinid whistles (Datta and Sturtivant 2002), detect leopard seal calls (Klinck et al. 2008), and recognize killer whale calls (Brown and Smaragdis 2009). HMMs have been successfully applied to terrestrial bioacoustics as well (Adi et al. 2010; Clemins et al. 2005; Kéç-Kogan and Margoliash 1998).

15.4.6 Unsupervised Learning

Unsupervised learners, which typically take the form of clustering algorithms, attempt to discover the structure of data. Examples of this include Kohonen's self-organizing map, the *k*-means algorithm, Gaussian mixture models, and adaptive resonance theory networks. These may all be thought of as ways of clustering data. Kohonen's self-organizing maps cluster high-dimensional data on to a two (or at least low)-dimensional grid (Hastie et al. 2001). The *k*-means algorithm and GMMs, both mentioned above, can also be thought of as unsupervised learners when they are used to discover unlabeled clusters. One criticism of both algorithms is that they assume the number of clusters *a priori*. An alternative to this is adaptive resonance theory (ART) networks (Carpenter et al. 1991; Grossberg 1988) where clusters are constructed dynamically. ART networks consider the similarity between an input feature vector and cluster centers. If the feature vector is close enough to an existing cluster as determined by a threshold mechanism called vigilance, it is assigned to that cluster; otherwise a new cluster may be formed. Deecke and Janik (2006) have used a variation of the ART algorithm where the similarity was computed using dynamic time warping. They were able to successfully cluster signature whistles of bottlenose dolphins as well as killer whale calls.

15.4.7 Evaluating Classifier Performance

Data for a classifier should always be separated into at least training and validation sets. Due to the possibility of overfitting, classification of training data does not give a reliable indication of how well the system will perform on future data. Most classifiers have some type of tunable parameters, and it is common to set these experimentally by examining how well the system performs on a validation set. One view

of parameter tuning is that it is in effect a form of training (on the validation data) and then the question arises as to whether or not the results are indicative of future field performance. As a consequence, whenever feasible, it is highly recommended to have a separate set of data called an evaluation set that is not tested until after the final models are created.

N-fold cross-validation or leave-one-out cross-validation (Duda et al. 2001) are frequently used to deal with limited amounts of data. *N*-fold cross-validation consists of dividing one's training and validation data into *N* partitions (folds). One selects most of the partitions (perhaps 60–70 %) as training data, and then uses the remaining data for validation. This process is repeated *N* times, each time moving one fold into the training data and another one out. Leave-one-out cross-validation, or jackknifing the data as it is sometimes called, refers to training a model with all training samples except one and then testing on the left out element. This process is repeated for every sample. With either method, the average error is reported.

A common extension of this is bootstrap evaluation (Hastie et al. 2001), which attempts to estimate the bias and variance of a classifier. In bootstrap evaluation, multiple random samples are drawn from the training data. For each sample, an equivalently sized training set is used by drawing with replacement (the same sample can be drawn multiple times). A classifier is constructed for each random sample, and then the mean is taken as with the previous techniques. An advantage to this method is that one can estimate the bias and variance from the error rate statistics.

If the goal is to detect a certain event such as a specific call, specific individual, or calls from a specific species, it is common to use some type of threshold to make "accept" or "reject" decisions. Varying this threshold will result in changes to the false-positive and missed-call rates. It is common to plot how these two types of error vary with respect to threshold, and receiver operating characteristic (ROC) curves are a common type of such a plot (Swets 1964). One must have a set of scores for the calls of interest, and a separate set of scores for other calls that could be mistakenly detected. Figure 15.10 shows a sample ROC curve; the horizontal axis shows the false positive (or false alarm) rate and the vertical axis shows the true positive rate. Each point on the curve shows the two types of error rate for a specific threshold, although the threshold values cannot be inferred from the plot. Given the data used to create the ROC curve, it possible to determine the threshold for a desired operating point such as 90 % true positives and 8 % false positives.

An alternative to the ROC is the detection error tradeoff (DET) curve proposed by Martin et al. (1997). The DET curve has two major differences from the ROC curve. Rather than plotting on the vertical axis the rate at which calls are detected, the rate of missed calls is plotted. Martin et al. argue that plotting error on both axes is more appropriate, and as a result of this better performance occurs on the lower left of the plot as opposed to the upper left. A second and more fundamental change is to assume that the score distributions for the calls of interest and other calls are each normally distributed. The axes are scaled to the deviates of normal distributions fitted to each type of score. When score distributions are normal, this will result in a straight line as opposed to a curve, but more importantly, the DET curve makes it easier to see the differences between classifier systems. Figure 15.11

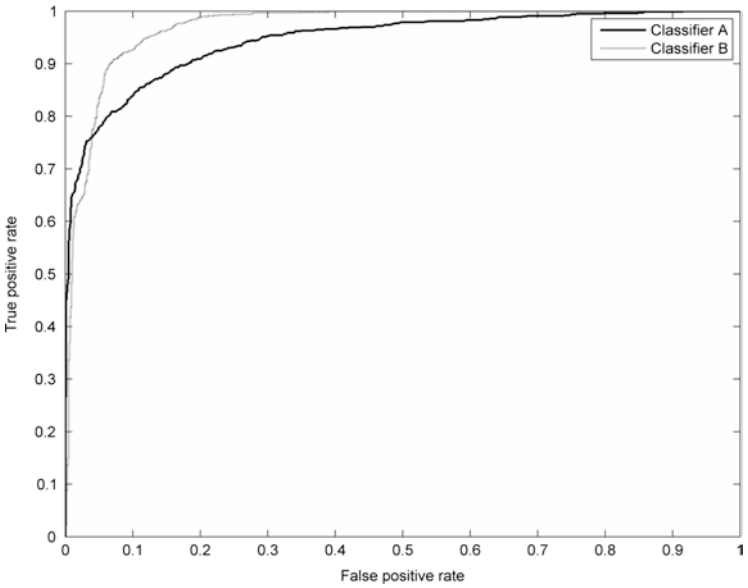

Fig. 15.10 An example of a receiver operating curve (ROC). The ROC shows the tradeoff in detection performance between correct detections and false positives as the decision threshold varies. Performance is better when the curve is closer to the *top left* of the plot. The performance of two hypothetical classifiers, where Classifier B outperforms Classifier A for most threshold values, is shown. See also Fig. 15.11

reports results for the same hypothetical classifiers shown in Fig. 15.10, but provides better separation between the curves, making it easier to compare systems. The National Institute of Standards and Technology provides software for producing DET plots in both Matlab and gnuplot (NIST 2010).

Two other performance measures widely used for evaluating detectors are precision and recall (Table 15.1). *Precision* is the fraction of all detections that are correct (true) detections. *Recall* is the fraction of all true instances that are successfully detected; it is equal to one minus the false-negative rate.

When considering any of the aforementioned techniques for acoustic data, one should be very aware that it is easier to recognize calls collected from similar environments than calls whose environments differ. As an example, one would expect better performance when the bathymetry and sea state are similar. Changes in environment can have serious impact on the feature set, and one may find that a classifier has learned a specific environment rather than species or call. This problem is not unique to bioacoustics, and has its parallels in both speech processing (Huang et al. 2001) and music identification (Downie 2008). This is illustrated in Fig. 15.12, which shows the data of Fig. 15.7 comparing the first two cepstral features of Pacific white-sided dolphin and Risso's dolphin echolocation clicks with the addition of data from a second sighting of Risso's dolphins. In spite of compensating for differences

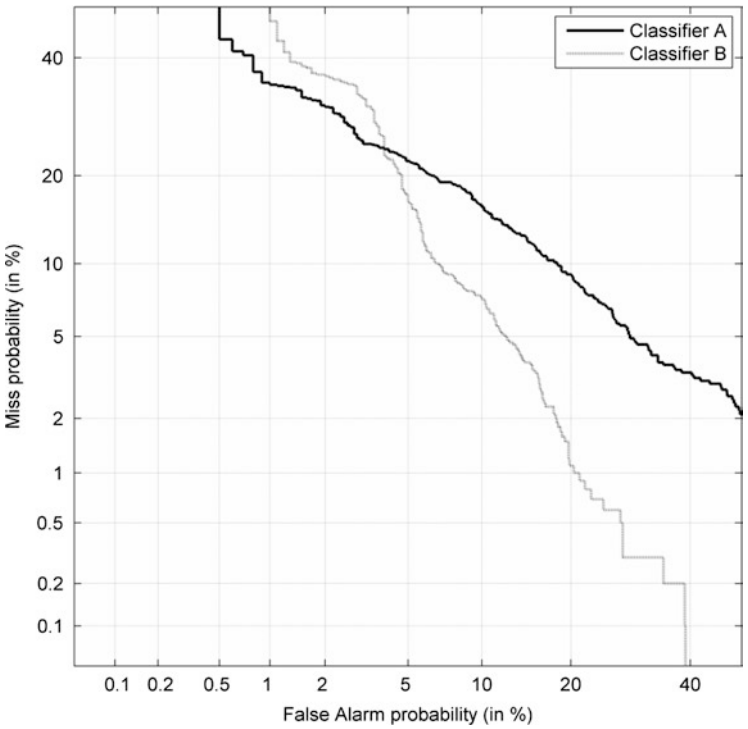

Fig. 15.11 The detection error tradeoff (DET) curve. DET curves assume that scores are distributed normally and plot normal deviates. This plot summarizes performance data for the same hypothetical classifiers shown Fig. 15.10, but highlights differences between the two classifiers

between collection systems by subtraction of the transfer function from the spectra, the distribution of the second sighting of Risso's dolphins has shifted.

As a consequence, the authors recommend that regardless of the evaluation method, all data from the same sighting should be either entirely in the training data or entirely in the test data. Splitting similar data across the train/test boundary is quite likely to improve results for the dataset being tested, but is unlikely to give one a good estimate of field performance (i.e., it will have poor generalization).

15.5 Localization

Passive acoustic localization refers to the use of acoustic signals to estimate the position of vocalizing marine life. Localization methods are useful for monitoring efforts as well as in studies of behavior, distribution, abundance, and acoustics. Various methods have been developed for different applications according to the number and configuration of hydrophones, the sound signal characteristics

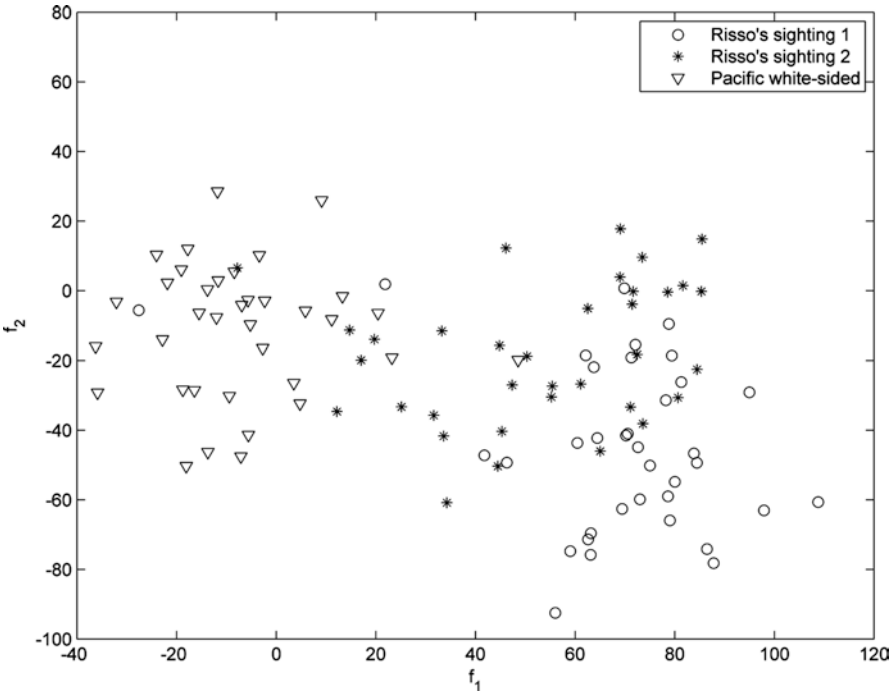

Fig. 15.12 Comparison of effects from different field collection situations. The first two cepstral features for the same dataset shown in Figure are plotted along with features from a second sighting of Risso's dolphins. Note how the distribution of Risso's dolphin features from the second sighting is less well separated from the Pacific white-sided dolphin. Shifts in features between collection situations are common and can arise from multiple sources (see text). The authors do not recommend splitting data from the same sighting when selecting training and test partitions

(duration, bandwidth, directivity, and so on), the operational requirements (such as required accuracy and precision of position estimates and computational efficiency), and the acoustic environment through which the signal propagates.

Most passive acoustic localization methods rely on travel times between the source and receivers. Unfortunately, the time at which an animal makes a call is unknown so it is not possible to measure travel time directly. Instead, most methods use the difference in arrival times between two or more receivers, since these times are independent of the time at which a call is generated. Such methods usually require a system with two or more hydrophones, called a hydrophone array. Since locations are calculated from arrival times, hydrophones must be synchronized and their positions known (often a nontrivial matter). Array processing falls into two broad categories depending on hydrophone spacing and distances over which animals are to be localized, either a compact or a widely spaced array.

This section describes localization methods descriptively; for details on related equations and calculations, the reader is referred to the references listed (which is by no means a comprehensive list). A useful overview of some localization methods, complete with equations, derivations, and Matlab™ code, is given in Zimmer (2011).

15.5.1 Compact Hydrophone Arrays

If the separation of the hydrophones is small compared to the distance of the sound source from the array, the incident sound can be approximated as a plane wave. In this case, beamforming is used to estimate the angle to the source (Johnson and Dudgeon 1993). In the simplest version, called time-domain beamforming, the arrival delay at each hydrophone is calculated for each possible arrival angle. The inverse of these delays is applied to each hydrophone signal and the resulting signals are summed. When the array is "steered" at the correct angle (by choosing the angle of the source), the delayed signals from all hydrophones coincide to give one loud combined signal; at other angles, the signals from the source interfere instead of coinciding, which results in a weaker signal. More hydrophones result in higher array gain (better signal-to-noise ratio for signals in the steered direction) and higher degrees of directionality.

A common configuration for beamforming is a linear array of hydrophones (Leaper et al. 1992; Sayigh et al. 1993; Miller and Tyack 1998). Only the angle of the source relative to array axis is obtained, which results in a cone of source position ambiguity—a 3D rotation about the axis of a line defined by the angle (Fig. 15.13). In many situations two-dimensional solutions are adequate, and the ambiguity cone is reduced to a curve (given by the intersection of the cone with a plane). This results in a left/right ambiguity for a horizontal array. Situations in which 2D solutions are adequate include when the water depth is small compared to the distance involved or when animals vocalize at predictable depths, such as near the surface. Position ambiguities

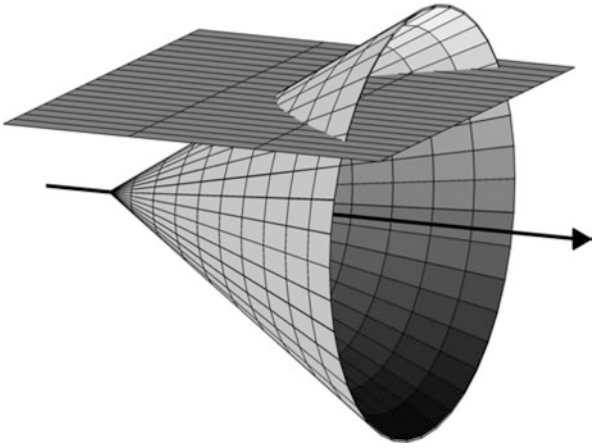

Fig. 15.13 Source position ambiguity cone for a horizontal linear array. Ambiguity can be reduced to a curve if the source depth is known by intersection with a plane corresponding to the source depth. The elements of the array lie along the horizontal axis represented by the *arrow*

can also be reduced by using more than one array (e.g., Watkins et al. 2000), provided that the spacing between arrays is wide enough to give sufficient bearing differences. Another approach uses time-motion analysis of the changes in estimated source angles as the array is towed (Leaper et al. 1992; Barlow and Taylor 2005). This method requires that vessel speed be much greater than the speed of the vocalizing animal, that the animal vocalize continuously for several minutes, and that individuals vocalizing simultaneously can be distinguished. In some cases, additional information can also be used to resolve ambiguities (for example, see multipath processing below).

Compact arrays can be built in many different configurations. For example, Clark (1980) used a compact 3-element planar array to estimate the unambiguous bearing to southern right whale calls. Compact planar arrays have also been successfully used for echolocation research (e.g., Rasmussen et al. 2002; Au et al. 2004). Planar arrays remove the bearing ambiguity of linear arrays; the ambiguity surface is the intersection of two cones, one along each axis of the array. However, without further information one cannot resolve which side of the plane a source is on. Three-dimensional arrays can resolve this array plane ambiguity (Wiggins et al. 2012; Zimmer 2013), and are becoming increasingly popular for this reason.

Optimally for continuous wave signals (that is, for long duration signals of single frequency and constant amplitude), hydrophone spacing must be less than half a wavelength and the largest dimension of the array, called the aperture, must be at least several wavelengths. For signals that are not continuous wave (e.g., impulsive or frequency-modulated calls), which is the case many marine mammal vocalizations, wider receiver spacing can often be used. In these cases, the receiver spacing should be close enough (usually within tens of meters) to ensure signal coherence across the receivers and beampatterns for the array should be calculated so performance is clearly understood (as shown for example in Zimmer 2011). In these cases it is often possible, and more computationally efficient, to use time-difference-of-arrival methods (see Sect. 15.5.2) with calculations simplified by the plane wave assumption.

15.5.2 Widely Spaced Hydrophone Arrays

Different methods are used when the source-receiver spacing is less than the spacing of the hydrophones, in which case the plane wave assumption is violated. The signal reaches two spatially separated receivers at different times because of different propagation path lengths from the source to the receivers. The difference in arrival time is called the time difference of arrival, or TDOA. TDOA methods are generally most accurate for sources near the center of the array, with decreasing accuracy as a source moves away from the array.

For two known receiver positions and a given TDOA, the locus of possible source locations in three dimensions is a hyperboloid. A third receiver provides another TDOA measurement, which defines a second hyperboloid (the third hydrophone actually adds two TDOAs but only one new TDOA is unique). A curve of possible source locations is defined by the intersection of these two hyperboloids. A fourth receiver defines a third hyperboloid, which intersects the curve at one or two points,

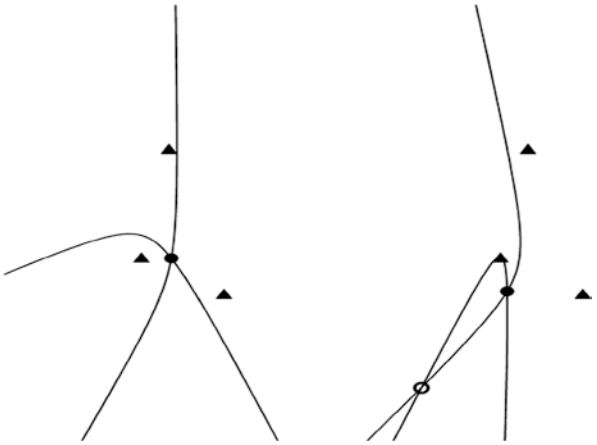

Fig. 15.14 Unambiguous 2D localization is possible with 3 hydrophones (triangles) in one case (left) but not another (right). True/false sources are shown with filled/open circles.

depending on the receiver geometry and animal position. In general, a fifth receiver is required to localize in three dimensions without ambiguity (Tyrrell 1964; Spiesberger 2001). However, for a given receiver configuration, there are usually large spatial regions for which only four receivers are sufficient for 3D localization (Spiesberger 2001). In these regions, either the source/receiver geometry results in a single point of intersection, or physical constraints (e.g., the seafloor or land) eliminate one of the source position ambiguities. On the other hand, even five hydrophones can give infinitely many possible source locations in some degenerate configurations. As discussed for compact arrays, 2D solutions are often sufficient, in which case the hyperboloids are reduced to hyperbolas and only four hydrophones are required to locate the source (and three hydrophones suffice in some regions). Figure 15.14 shows a 2D case for which three hydrophones are sufficient for one whale position but not for another.

Assuming that sound speed is spatially homogeneous, the problem of finding the point of intersection of the hyperboloids (or the closest such point if intersection is imperfect) can be expressed as a system of linear equations. For a well-defined problem (not underdetermined/overdetermined by too few/many receivers), a closed form solution to this system gives the source location (e.g., Schmidt 1972; Watkins and Schevill 1971). For overdetermined systems, a least-squares approach can be used to give the best source position (Spiesberger and Fristrup 1990; Wahlberg et al. 2001); the extra hydrophones reduce the error in the position estimate.

15.5.3 Nonhomogeneous Sound Speed

Homogeneous sound speed assumptions can result in poor location estimates when long distances or shallow water are involved (Chapman 2004). For widely spaced arrays, nonhomogeneous sound speeds can be accounted for by using nonlinear methods that incorporate differences in sound speed to construct probability density

functions for source position. One approach assumes a different sound speed between the source and each of the receivers and solves the resulting nonlinear system (Spiesberger and Wahlberg 2002). Another approach, sometimes referred to as model-based tracking, allows the sound speed to vary with depth but not X-Y position (Tiemann et al. 2004; Thode 2005). A source is localized by finding the source position that gives predicted arrival times that best match the measured arrival times. Arrival time predictions are made using a sound propagation model, which in turn uses information about the environment including sound speed profiles and bathymetry.

15.5.4 Establishing Time-of-Arrival Differences

For relatively loud and/or impulsive (sharply peaked) signals in small datasets (and with very patient observers), arrival time difference can be estimated manually through visual inspection of raw or filtered waveforms or spectrograms. Since this is an extremely tedious process that can be especially difficult in noisy conditions, automated techniques to establish TDOAs are commonly used.

One way to automatically establish TDOAs is to use a "detect and associate calls" approach. A detection method (see Sect. 15.4) is used to find all calls of interest on all hydrophones. The same call (or call sequence) is associated over all hydrophones—that is, each call is associated with its arrivals on the multiple hydrophones—and arrival times of associated calls establish TDOAs. Call association can be a simple task for a single animal or when calling rates are low, such that each call is easily identified across multiple hydrophones. For more difficult cases with multiple animals with high calling rates, one option is to create histograms of TDOAs from all possible associations over a time period long enough to contain multiple calls from an individual animal. Since TDOAs vary slowly with animal movement, correctly associated calls will result in histogram peaks (e.g., Morrissey et al. 2006). Another approach separates sources before association, for example by tracking slowly varying features such as amplitude, frequency, inter-call intervals, and so on (e.g., Clark 1989). This "detect and associate" method requires that calls are sufficiently stereotyped for detection but variable enough to distinguish individual calls.

A commonly used method used for establishing TDOAs that does not require stereotyped calls is known as cross-correlation (Helstrom 1975). The TDOA estimate is the time-lag that maximizes the cross-correlation between received signals from two hydrophones. Both filtered waveforms and spectrograms of the recorded signals have been used for cross-correlation (Altes 1980; Clark et al. 1986; Spiesberger and Fristrup 1990). The cross-correlator provides gain in the signal-to-noise ratio resulting in greater ranges over which an animal can be localized. Since cross-correlation assumes that the received signal at each hydrophone is the same except for a time lag, there are cases in which it does not perform well. Such cases include highly directional call components, complicated propagation conditions, or animals that move quickly while vocalizing so that Doppler effects become important. Multiple animals can be localized by picking multiple peaks in the

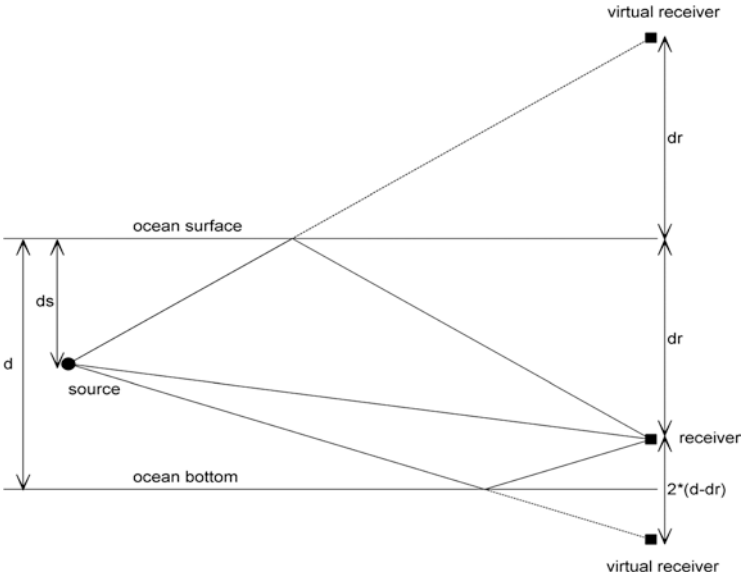

Fig. 15.15 Virtual receiver arrivals (*dotted lines*) corresponding to multipath arrivals (*solid lines*) for a flat bottom

cross-correlation function, although care is required to avoid confusion from multipath arrivals (Spiesberger 2000) and mis-association between animals (Baggenstoss 2011). Some multiple animal localization methods are designed to handle spurious/incorrect TDOAs to ease this requirement (Baggenstoss 2011; Nosal 2013).

15.5.5 Reflection Methods

In cases when multipath arrivals exist and can be separated, reflected paths can be used to help localize a sound source. To use reflections, the TDOA method can be modified by adding a virtual hydrophone that corresponds to each reflection (Fig. 15.15). The time delay between the direct-path arrival and the reflection arrival is proportional to the additional distance present in the reflection path compared to the direct path. Note that water-borne acoustic signals that reflect off the water's surface are inverted, so methods that use cross-correlation with surface reflections need to reverse the sign of the correlation result. Reflections can be used to resolve position ambiguities and improve the accuracy of estimated source positions (Wahlberg et al. 2001; Thode et al. 2002; Zimmer et al. 2003), or to localize a source with nonsynchronized hydrophones (Nosal and Frazer 2006). They can also be used to reduce the number of hydrophones needed for localization; using multipath arrivals, a single hydrophone can be used to estimate the range and depth of a calling animal (Cato 1998; Aubauer et al. 2000; Širović et al. 2007). If bathymetry varies

with azimuth, an animal can be located in 3D using a single hydrophone if enough reflections can be extracted (Tiemann et al. 2007).

Reflection methods cannot be used for tonal long-duration signals in which various reflected arrivals cannot be separated. Even for short-duration signals it is not always possible to distinguish reflections. For example, very shallow vocalizations (or very shallow hydrophones) will result in direct and surface-reflected arrival times that are nearly identical. For highly directional vocalizations, such as clicks from many species of odontocetes, there might be insufficient off-axis energy to give a reflected arrival, or even a direct arrival when the reflection is strong.

15.5.6 Error Estimates

Just as important as finding source positions is understanding the errors in the estimates. Most errors in position estimate stem from uncertainty in receiver position, TDOA estimates, and sound speed. The most direct way to quantify error is to localize sources with known position. A controlled source can be used for this purpose (e.g., Watkins and Schevill 1972; Janik et al. 2000; Clark and Ellison 2000), or positions can be verified visually (Frankel et al. 1995; Noad and Cato 2007; Tiemann et al. 2006). This direct approach can be difficult to apply and generalize since resulting errors are specific to the call type, environment, and source/receiver geometry.

For practical reasons, error is usually estimated theoretically. Linear error propagation is a simple and powerful approach with much literature devoted to it (Taylor 1997; Watkins and Schevill 1971; Spiesberger and Fristrup 1990; Wahlberg et al. 2001). However, because source location is not a linear function of the model inputs, linear propagation methods can significantly overestimate error bounds and nonlinear methods can give more accurate error bounds (in addition to more accurate position estimates). For methods that construct probability density functions to localize a source, confidence regions can be defined by curves/surfaces of constant probability density (Clark and Ellison 2000; Spiesberger and Wahlberg 2002). Error can also be estimated through sensitivity studies that use a simulated source localized in perturbed environments (Tiemann et al. 2004; Thode 2005). Ideally this last approach would use a scheme such as Monte Carlo to repeat localizations for different perturbations of the environment to account for all uncertainties and their distributions.

In addition to practical issues, error is an important consideration when designing an array. Error maps can be used to optimize the hydrophone configuration and placement so that errors are minimized in the areas where sources are to be localized. For example, error analysis for a linear array reveals that angle estimates are most accurate for sources perpendicular to the array axis and the least accurate for sources in line with the array. An example of an error map for a widely spaced array is given in Fig. 15.16; sound sources within the array can be localized quite accurately, but accuracy decreases rapidly as a source moves away from the array.

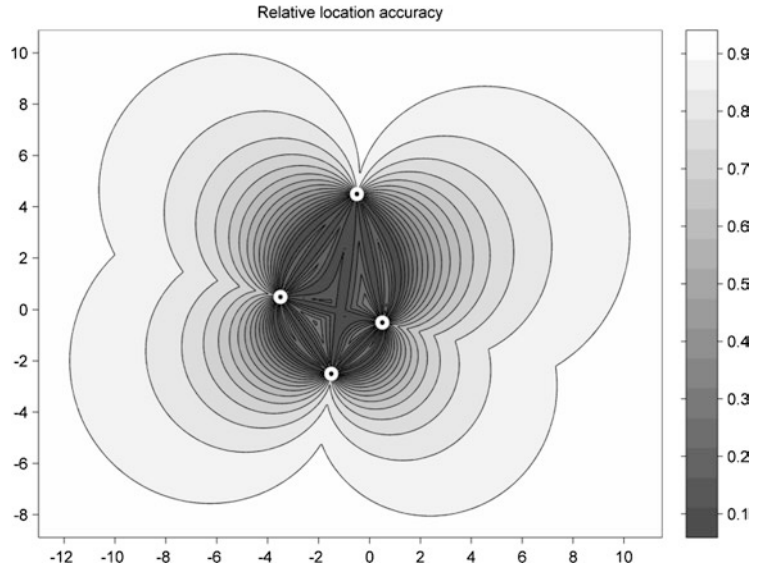

Fig. 15.16 Relative accuracy of locations calculated using the 2-D TDOA method for a given hydrophone configuration (*black dots* in *white circles*). Accuracy is good at the center of the array but falls away with distance, especially at the corners. Values represent summed location error per unit change in position

15.5.7 Other Approaches

Most localization methods rely on arrival times because they are quite robust and so can give accurate position estimates. However, other information about the position of a calling animal is available in a received signal and can be used to obtain or improve position estimates. For example, a directional hydrophone can provide a rough bearing estimate (Whitehead and Gordon 1986). If propagation effects are carefully accounted for and the call is not highly directional, differences in received levels on two or more omnidirectional sensors can be used to locate an animal (Cato 1998; Frank and Ferris 2011). Mode dispersion can also be used to estimate the range of low-frequency animal calls (McDonald and Moore 2002; Wiggins et al. 2004; Newhall et al. 2012). Matched-field processing (MFP) (Tolstoy 1993; Thode et al. 2000) finds the source position that predicts the acoustic field most similar to the measured field (note that the TDOA method can be thought of as MFP in which the only part of the field that is matched is arrival time).

Although simpler methods are often adequate, more sophisticated techniques and sensors can be used to improve localization capabilities. For example, more accurate position and error estimates can be obtained when localization is treated as a joint inversion problem for source position, receiver position, sound speed, and/or other relevant parameters (e.g., seafloor characteristics and sea state) (Tarantola 1987; Spiesberger and Fristrup 1990; Thode et al. 2000; Rideout et al. 2013).

Another promising development is in sensors (e.g., vector sensors or DIFAR buoys) that measure particle velocity in addition to pressure, allowing arrival direction to be estimated using a single sensor (McDonald 2004; Greene et al. 2004; Thode et al. 2010). Hopefully such powerful approaches will become more widely accessible as computing resources and sensors become less expensive and as efforts continue to develop improved localization methods.

15.6 Software

The availability and capability of software packages vary quickly over time, and consequently only a brief survey of available tools will be given. Discussion is divided into tools designed with bioacousticians in mind and those that are more general pattern recognition toolsets. Websites are just as transient if not more so, and if a URL given below does not function, a web search engine is likely to reveal the new site if the package still exists.

The three most common freely available software packages in marine mammal bioacoustics are Ishmael (Mellinger 2001), PAMGUARD (Gillespie et al. 2008), and XBAT (Figueroa and Robbins 2007). Available software packages for bioacoustic data analysis can be categorized into two groups: real-time and post-processing software packages. Real-time software tools allow users to record acoustic data and to run detection, classification, and localization algorithms in real time on incoming data streams. Ishmael and PAMGUARD fall into this category and are commonly used for ship-based passive acoustic surveys for which real-time capabilities are essential. XBAT is a post-processing software package developed to analyze field recordings in the lab and does not at this point provide recording capabilities.

Ishmael, PAMGUARD, and XBAT allow users to explore data in the time (waveform) and frequency (spectrogram) domains and are capable of detecting/classifying and localizing sounds of interest. All three programs are controlled via a graphical user interface which provides easy access to the main functions of the program. However, there are some significant differences in functionality of each software package elucidated in the following paragraphs which provide a brief introduction to the capabilities and goals of each package. For a more detailed description of the software packages and their modules, refer to the corresponding publications, websites, and user's manuals.

15.6.1 Ishmael (http://www.bioacoustics.us/ishmael.html)

The current version of Ishmael can be operated stand-alone on Windows™, Linux, and Macintosh platforms (the latter two under the WINE wrapper). Ishmael is capable of recording sounds and running detection and localization algorithms on incoming data streams. It handles a variety of data acquisition hardware and is well

suited for real-time applications such as ship-based passive-acoustics surveys or analysis of long-term data sets from fixed hydrophones. Six detection and four localization methods are available in Ishmael. Detection methods are based on matched filtering, spectrogram correlation, energy summation, frequency contour detection (whistles and moans), the Teager–Kaiser energy operator (clicks), and characteristic repetition patterns of sounds. In recent versions of Ishmael, multiple views and multiple detectors may be run in parallel, allowing detection using multiple detection methods or parameters, or detection of multiple call types. Localization methods are based on phone-pair bearing estimation, hyperbolic position estimation, beamforming, and crossed bearings from two hydrophone pairs. Ishmael can also be operated in post-processing mode and batch run functionality allows a user to run detection algorithms over large data sets.

15.6.2 PAMGUARD (http://www.pamguard.org)

PAMGUARD is a Java™ based program which can be run on all major operating systems (Windows, Mac OS, and Linux). PAMGUARD was originally developed for ship-based passive-acoustics surveys, though it is also useful for post-processing data in files. A communications interface allows a user to access GPS data streams and to visualize ship tracks as well as locations of acoustic detections via a mapping component. PAMGUARD can interface to a wide variety of hardware to capture sound. It features five detection, one classification, and three localization methods. The available detection algorithms are based on matched filtering, spectrogram correlation, energy summation, frequency contour detection (whistles and moans), and the Teager–Kaiser energy operator (clicks); multiple instances of detectors can be run in parallel to try different detection methods and parameters, or to search for different call types. The built-in classifier can be used for real-time whistle classification. Available localization methods are phone-pair bearing estimation and hyperbolic position estimation. PAMGUARD is a modular program which can be extended by any Java™ programmer. Detailed information on how to do this can be found on the PAMGUARD website and in the user's manual.

15.6.3 XBAT (http://www.xbat.org)

XBAT (Figueroa and Robbins 2007) is a post-processing software package to analyze field recordings in the lab. In contrast to Ishmael and PAMGUARD, XBAT is not a stand-alone application: Matlab™ is necessary to be able to run the software. XBAT features an extensive input module which can handle a large selection of file formats (including compression codecs such as mp3, ogg-vorbis, and flac). The software can be configured to load many consecutive files as a file stream, which is useful to display long-term spectrograms and for visual exploration of acoustic data.

The main detection module of XBAT, based on spectrogram correlation combined with nearest-neighbor search, is easy to use. The user marks a sound of interest in the spectrogram and XBAT uses this template to search for similar sounds in the data set. The spectrogram correlation module can handle several templates at the same time, which allows a user to search for different sounds in parallel. Also templates for confounding sounds can be configured to reduce the number of false positive detections. Sounds of interest recorded on several channels can be localized by hyperbolic position estimation. XBAT is a modular software package which can be extended by any Matlab™ programmers. However this is not trivial, as there is very little documentation available on how to do this.

15.6.4 Additional Software Packages

A number of companies, institutions, and individuals offer commercially or freely available software packages designed for bioacoustic research or general scientific signal analyses. However the description of these software tools is beyond the scope of this chapter. For more information on additional software tools, please visit the "About Bioacoustics" page at http://tcabasa.org.

15.6.5 General Pattern Recognition Software

For general pattern recognition software, one can separate the types of available software into complete packages versus stand-alone libraries that offer one or more classifiers to be integrated. WEKA and the hidden Markov model toolkit (HTK) offer complete recognition systems. WEKA (Hall et al. 2009) is a graphically oriented system designed to provide an interface for classification and regression. It provides an interface for a wide variety of learning algorithms. In contrast, HTK (Young et al. 2006) was developed for speech processing and is widely used in that community. Unlike WEKA, the focus is entirely on functionality, and commands and errors can be cryptic. It implements hidden Markov models, Gaussian mixture models, and *k*-means clustering, and requires a large learning curve. Finally, the R language (R Development Core Team 2009) is an open-source language developed for statistical analysis which has a large number of classifiers as add-on packages.

Other systems provide libraries that can be linked to programming languages such as python™, Java™, and Matlab® and are candidates for practitioners with good programming skills. Examples include JBoost (a boosting library; http://jboost.sourceforge.net) and the Torch machine learning library (http://www.torch.ch and http://torch5.sourceforge.net).

15.7 Future Directions

It is the authors' opinion that the greatest gains to be made lie in the realm of feature extraction. Whether working with frequency contours or echolocation clicks, feature extraction is difficult. Most systems working with frequency contours do not attempt to account for the shape of the contour, with notable exceptions of the dynamic time warping, matched filter, and spectrogram correlation methods (Deecke and Janik 2006; Mellinger and Clark 2000). Instead, they examine statistics of the whistle such as frequency maxima and number of inflection points, which do not capture the shape. When asking researchers examining spectrograms why a specific call should be associated with a species, pod, or call type, many will reply with something along the lines of "it just *looks* that way." Features that capture this type of shape information as well as those that are capable of handling nonlinear phenomena are likely to yield advances, but an alternative and perhaps better approach is to invest more attention into how the animals are likely to *perceive* calls.

In the study of echolocation clicks, features such as zero crossings, peak values and energy band ratios, and characteristics of spectral shape such as cepstra or spectral ridge regression parameters, are all commonly used features, but they fail to account for axis variation and high frequency falloff as distance increases. While some of this can be compensated for by classifiers that learn the patterns that occur, features that are more invariant under these conditions have the potential to produce significant advances in the field. As with the discussion of frequency contours, taking inspiration from perception is also likely to be fruitful.

While improved feature extraction appears to be the most promising direction for reducing classification error, ensemble methods have been a fruitful area of research in pattern recognition and bear brief mention. The principal idea is to create multiple models for each class. Bagging (Hastie et al. 2001) attempts to reduce errors by taking N bootstrap samples (sampling with replacement the same number of vectors as in the training sample) and creating a classifier for each one. The output of these classifiers are fused to create a single decision. Boosting (Freund and Schapire 1999) uses multiple classifiers, each of which is rather weak in that by itself it might perform only slightly better than chance. Rather than taking bootstrap samples as bagging does, each training vector is assigned an initially equal weight, and a weak classifier is created. The weights are adjusted to emphasize training samples that were misclassified by the weak model, and a new classifier is created. This process is iterated, and Freund describes this process as a means of focusing on the difficult cases (Yoav Freund, pers. comm., 2010). The final decision is made based upon a weighted average of all of the classifiers. Another popular ensemble technique that has been used in marine mammal acoustics (e.g., Henderson et al. 2011; Risch et al. 2013) is random forests, where multiple decision trees are formed from bootstrapped datasets and multiple trees vote (Hastie et al. 2009).

Another major challenge for passive-acoustic monitoring systems is the analysis of very large datasets. Due to the rapid development of digital audio technology and the increasing capacity of memory devices, it has become easier than ever to produce

very large long-term acoustic datasets that require considerable computation time to analyze. Parallel computing is a powerful tool to speed up the analysis of such large data sets. One approach to parallel computing uses multi-core processors (MCPs) within a single workstation. An easy way to make use of multi-core processors is to run several copies of an analysis program in parallel, with the operating systems automatically distributing the processes to all cores available. A more elegant way to benefit from a multi-core processor is to use software which can distribute computation tasks to all available cores, such as the parallel and distributed computing toolboxes for Matlab™ (Sharma and Martin 2009). A second approach to parallel computing is to use a graphics processing unit (GPU). A GPU, a collection of processors, typically handles computation for rendering computer graphics images. GPUs are powerful parallel computing devices, with hundreds or thousands of cores and many gigabytes of onboard memory. These can be used as general-purpose computers, or general purpose graphics processing units (GPGPUs). As with multi-core processors, the computations are distributed to all cores available; a Matlab toolbox for this is available. See Owens et al. (2007) for a more comprehensive description of GPGPUs.

Another approach is to use parallel computing on *clusters*—groups of computers linked to each other through a local area network (Thiruvathukal 2005). Setting up a parallel computing task on a cluster is more complex than execution on a single workstation. Data sets and a list of computation instructions are located on one or more servers within the local area network. The available processors repeatedly check the list of computation instructions for open jobs, download the respective data sets, conduct the analysis, and send the result back to the server(s). Since many data sets are transferred from server(s) to the processors, the throughput of the cluster may depend heavily on the bandwidth of the local area network. An example using a cluster to analyze bioacoustic data sets is given in Chap. 9 of this book.

A final cautionary word should be added about relying on parallel computing to achieve speed increases. Many times, the redesign of an inefficient algorithm can result in significant reductions in computing time. Most computer languages have profiling facilities that will let a user track how much time was spent in specific routines or even lines of code. Taking the time to determine where the "code bottlenecks" are and putting effort into redesign can offer significant improvements in performance that can either eliminate the need to invest time and capital in parallel architectures or at least provide even faster parallel implementations.

Acknowledgements This chapter was produced in part with funding from the Office of Naval Research for the "Advanced Detection, Classification, and Localization" project.

References

O. Adam, The use of the Hilbert-Huang transform to analyze transient signals emitted by sperm whales. Appl. Acoust. **67**, 1134–1143 (2006a)

O. Adam, Advantages of the Hilbert Huang transform for marine mammals signals analysis. J. Acoust. Soc. Am. **120**, 2965–2973 (2006b)

O. Adam, Segmentation of killer whale vocalizations using the Hilbert-Huang Transform. EURASIP J. Adv. Signal Proc. **10**, 2007–2016 (2008)
K. Adi, M.T. Johnson, T.S. Osiejuk, Acoustic censusing using automatic vocalization classification and identity recognition. J. Acoust. Soc. Am. **127**, 874–883 (2010)
R. Altes, Detection, estimation, and classification with spectrograms. J. Acoust. Soc. Am. **67**, 1232–1246 (1980)
W.W.L. Au, R.A. Kastelein, T. Rippe, N.M. Schooneman, Transmission beam pattern and echolocation signals of a harbor porpoise (*Phocoena phocoena*). J. Acoust. Soc. Am. **106**, 3699–3705 (1999)
W.W.L. Au, J.K.B. Ford, J.K. Horne, K.A. Newman Allman, Echolocation signals of free ranging killer whales (*Orcinus orca*) and modelling of foraging for chinook salmon (*Oncorhyncus tswawytscha*). J. Acoust. Soc. Am. **115**, 901–909 (2004)
R. Aubauer, M.O. Lammers, W.W.L. Au, One-hydrophone method for estimating distance and depth of phonating dolphins in shallow water. J. Acoust. Soc. Am. **107**, 2744–2749 (2000)
J. Barlow, B.L. Taylor, Estimates of sperm whale abundance in the northeastern temperate Pacific from a combined acoustic and visual survey. Mar. Mamm. Sci. **21**, 429–445 (2005)
P.M. Baggenstoss, An algorithm for the localization of multiple interfering sperm whales using multi-sensor time difference of arrival. J. Acoust. Soc. Am. **130**, 102–112 (2011)
M.F. Baumgartner, S.M.V. Parijs, F.W. Wenzel, C.J. Tremblay, H.C. Esch, A.M. Warde, Low frequency vocalizations attributed to sei whales (*Balaenoptera borealis*). J. Acoust. Soc. Am. **124**, 1339–1349 (2008)
B. Boashash, Estimating and interpreting the instantaneous frequency of a signal. IEEE Fund. Proc. IEEE **80**, 520–538 (1992)
F. Bénard, H. Glotin, Automatic indexing for content analysis of whale recordings and XML representation. EURASIP J. Adv. Signal Proc. **2010**, 695017 (2010). doi:10.1155/2010/695017
J.C. Brown, P.J.O. Miller, Automatic classification of killer whale vocalizations using dynamic time warping. J. Acoust. Soc. Am. **122**, 1201–1207 (2007)
J.C. Brown, P. Smaragdis, Hidden Markov and Gaussian mixture models for automatic call classification. JASA-EL **125**, EL222–EL224 (2009)
J.R. Buck, P.L. Tyack, A quantitative measure of similarity for *Tursiops truncatus* signature whistles. J. Acoust. Soc. Am. **94**, 2497–2506 (1993)
C.J.C. Burges, A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. **2**, 121–167 (1998)
G.A. Carpenter, S. Grossberg, D.B. Rosen, ART 2-A – an adaptive resonance algorithm for rapid category learning and recognition. Neural Netw. **4**, 493–504 (1991)
D.H. Cato, Simple methods of estimating source levels and locations of marine animal sounds. J. Acoust. Soc. Am. **104**, 1667–1678 (1998)
D.M.F. Chapman, You can't get there from here: shallow water sound propagation and whale localization. Can. Acoust. **32**, 167–171 (2004)
C.W. Clark, 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 (1980)
C.W. Clark, The use of bowhead call-tracks based on call characteristics as an independent means of determining tracking parameters. Sci. Rep. Intl. Whal. Commn. **39**, 111–113 (1989)
C.W. Clark, W.T. Ellison, K. Beeman, Acoustic tracking of migrating bowhead whales. Proc. IEEE Oceans **86**, 341–346 (1986)
C.W. Clark, W.T. Ellison, Calibration and comparison of acoustic location methods used during the spring migration of the bowhead whale, *Balaena mysticetus*, off Pt. Barrow, Alaska, 1984–1993. J. Acoust. Soc. Am. **107**, 3509–3517 (2000)
P.J. Clemins, M.T. Johnson, K.M. Leong, A. Savage, Automatic classification and speaker identification of African elephant (*Loxodonta africana*) vocalizations. J. Acoust. Soc. Am. **117**, 956–963 (2005)
N. Cristianini, J. Shawe-Taylor, *Support vector machines and other kernel-based learning methods* (Cambridge Univ. Press, Cambridge, 2000)

S. Datta, C. Sturtivant, Dolphin whistle classification for determining group identities. Signal Proc **82**, 127–327 (2002)
V.B. Deecke, J.K.B. Ford, P. Spong, 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 (1999)
V.B. Deecke, V.M. Janik, Automated categorization of bioacoustic signals: avoiding perceptual pitfalls. J. Acoust. Soc. Am. **199**, 645–653 (2006)
J.S. Downie, The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research. Acoust. Sci. Technol. **29**, 247–255 (2008)
R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification* (Wiley, New York, 2001)
C. Erbe, A.R. King, Automatic detection of marine mammals using information entropy. J. Acoust. Soc. Am. **124**, 2833–2840 (2008)
H. Figueroa, M. Robbins, XBAT: an open-source extensible platform for bioacoustic research and monitoring, in *International Academy for Nature Conservation*, ed. by K.-H. Frommolt, R. Bardeli, M. Calusen (BfN Skripten, Isle of Vilm, 2007), pp. 143–155
S.D. Frank, N. Ferris, Analysis and localization of blue whale vocalizations in the Solomon Sea using waveform amplitude data. J. Acoust. Soc. Am. **130**, 731–736 (2011)
A.S. Frankel, C.W. Clark, L.M. Herman, C.M. Gabriele, Spatial distribution, habitat utilization, and social interactions of humpback whales, *Megaptera novaeangliae*, off Hawaii, determined using acoustic and visual techniques. Can. J. Zool. **73**, 1134–1146 (1995)
Y. Freund, R.E. Schapire, A short introduction to boosting (Japanese, English version available at http://cseweb.ucsd.edu/~yfreund). J. Jpn Soc. Artif. Int. **14**, 771–780 (1999)
K.M. Fristrup, *Characterizing Acoustic Features of Marine Animal Sounds. Technical Report WHOI-92-04* (Woods Hole Oceanographic Inst., Woods Hole, MA, 1992)
C. Gervaise, A. Barazzutti, S. Busson, Y. Simard, N. Roy, Automatic detection of bioacoustics impulses based on kurtosis under weak signal to noise ratio. Appl. Acoust. **71**, 1020–1026 (2010). doi:10.1016/j.apacoust.2010.05.009
D. Gillespie, An acoustic survey for sperm whales in the Southern Ocean Sanctuary conducted from the RSV *Aurora Australis*. Rep. Intl. Whal. Commn. **47**, 897–907 (1997)
D. Gillespie, Detection and classification of right whale calls using an edge detector operating on a smoothed spectrogram. Can. Acoust. **32**, 39–47 (2004)
D. Gillespie, R. Leaper, Detection of sperm whale (*Physeter macrocephalus*) clicks and discrimination of individual vocalisations. Eur. Res. Cetaceans **10**, 87–91 (1996)
D. Gillespie, J. Gordon, R. McHugh, D. McLaren, D.K. Mellinger, P. Redmond, A. Thode, P. Trinder, X.-Y. Deng, PAMGUARD: semiautomated, open source software for real-time acoustic detection and localisation of cetaceans. Proc. Inst. Acoust. **30**, 9 (2008)
D. Gillespie, M. Caillat, J. Gordon, P. White, Automatic detection and classification of odontocete whistlesa). J. Acoust. Soc. Am. **134**, 2427–2437 (2013)
C.R. Greene, M.W. McLennan, R.G. Norman, T.L. McDonald, R.S. Jakubczak, W.J. Richardson, Directional frequency and recording (DIFAR) sensors in seafloor recorders to locate calling bowhead whales during their fall migration. J. Acoust. Soc. Am. **116**, 799–813 (2004)
S. Grossberg, *Neural Networks and Natural Intelligence* (MIT Press, Cambridge, 1988)
B.M. Gur, C. Niezrecki, Autocorrelation based denoising of manatee vocalizations using the undecimated discrete wavelet transform. J. Acoust. Soc. Am. **122**, 188–199 (2007)
M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update. SIGKDD Explorations **11**, 10–18 (2009)
G.D. Hastie, T.R. Barton, K. Grellier, P.S. Hammond, R.J. Swift, P.M. Thompson, B. Wilson, Distribution of small cetaceans within a candidate Special Area of Conservation; implications for management. J. Cetacean Res. Manag. **5**, 261–266 (2003)
T. Hastie, R. Tibshirani, J.H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 1st edn. (Springer, New York, NY, 2001)
T. Hastie, R. Tibshirani, J.H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn. (Springer, New York, NY, 2009)

E.E. Henderson, J.A. Hildebrand, M.H. Smith, Classification of behavior using vocalizations of Pacific white-sided dolphins (*Lagenorhynchus obliquidens*). J. Acoust. Soc. Am. **130**, 557–567 (2011)
C.W. Helstrom, *Statistical Theory of Signal Detection* (Pergamon, New York, NY, 1975)
D.S. Houser, D.A. Helweg, P.W. Moore, Classification of dolphin echolocation clicks by energy and frequency distributions. J. Acoust. Soc. Am. **106**, 1579–1585 (1999)
X. Huang, A. Acero, H.W. Hon, *Spoken Language Processing* (Prentice Hall, Upper Saddle River, NJ, 2001)
C. Ioana, C. Gervaise, Y. Stéphan, J.I. Mars, Analysis of underwater mammal vocalizations using time-frequency-phase tracker. Appl. Acoust. **71**, 1070–1080 (2010)
V.M. Janik, S.M. Van Parijs, P.M. Thompson, A two-dimensional acoustic localization system for marine mammals. Mar. Mamm. Sci. **16**, 437–447 (2000)
S.M. Jarvis, N.A. DiMarzio, R.P. Morrissey, D.J. Moretti, A novel multi-class support vector machine classifier for automated classification of beaked whales and other small odontocetes. Can. Acoust. **36**, 34–40 (2008)
A.T. Johansson, P.R. White, An adaptive filter-based method for robust, automatic detection and frequency estimation of whistles. J. Acoust. Soc. Am. **130**, 893–903 (2011)
D.H. Johnson, D.E. Dudgeon, *Array Signal Processing: Concepts and Techniques* (Prentice Hall, Englewood Cliffs, NJ, 1993)
J.F. Kaiser, On a simple algorithm to calculate the "Energy" of a signal. *Proc. IEEE Intl. Conf. Acoust., Speech, Sig. Process.*, Albuquerque, 381–384 (1990)
V. Kandia, Y. Stylianou, Detection of sperm whale clicks based on the Teager-Kaiser energy operator. Appl. Acoust. **67**, 1144–1163 (2006)
V. Kandia, Y. Stylianou, Detection of clicks based on group delay. Can. Acoust. **36**, 48–54 (2008)
J.A. Kéç-Kogan, D. Margoliash, 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 (1998)
A. Kirshenbaum, M.A. Roch, An image processing based paradigm for the extraction of tonal sounds in cetacean communications. J. Acoust. Soc. Am. **134**, 4435 (2013)
H. Klinck, L. Kindermann, O. Boebel, Detection of leopard seal (*Hydrurga leptonyx*) vocalizations using the envelope-spectrogram technique (tEST) in combination with a hidden Markov model. Can. Acoust. **36**, 118–124 (2008)
H. Klinck, D.K. Mellinger, The energy ratio mapping algorithm (ERMA): a tool to improve the energy-based detection of odontocete clicks. J. Acoust. Soc. Am. **129**, 1807–1812 (2011)
P. Kovesi, Phase-preserving denoising of images, in *Proc. Australian Patt. Rec. Soc. Conf: DICTA'99,* Perth, WA, Australia, 1999, p. 212–217
R. Leaper, O. Chappell, J. Gordon, The development of practical techniques for surveying sperm whale populations acoustically. Rep. Intl. Whal. Commn. **42**, 549–560 (1992)
R.P. Lippmann, Pattern classification using neural networks. IEEE Commun. Mag. **27**, 47–64 (1989)
S.K. Madhusudhana, M.A. Roch, E.M. Oleson, M.S. Soldevilla, J.A. Hildebrand, Blue whale B and D call classification using a frequency domain based robust contour extractor. *Proc. OCEANS '09*, Bremen, Germany, 2009, p 1–7
A. Mallawaarachchi, S.H. Ong, M. Chitre, E. Taylor, Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles. J. Acoust. Soc. Am. **124**, 1159–1170 (2008)
C.D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval* (Cambridge University Press, New York, NY, 2008)
A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET curve in assessment of detection task performance, in *Proc. Eurospeech 4*, Rhodes, Greece, 1997, p. 1895-1898.
D. Mathias, A. Thode, S.B. Blackwell, *Computer-aided classification of bowhead whale call categories for mitigation monitoring. Proc IEEE Passive, Hyères, France* (IEEE, Hyeres, 2008), p. 6
M.A. McDonald, DIFAR hydrophone usage in whale research. Can. Acoust. **32**, 155–160 (2004)

M.A. McDonald, S.E. Moore, Calls recorded from North Pacific right whales (*Eubalaena japonica*) in the eastern Bering Sea. J. Cetacean Res. Manag. **4**, 261–266 (2002)
M.A. McDonald, J.A. Hildebrand, M. Sarah, Worldwide decline in tonal frequencies of blue whale songs. Endanger Species Res. **9**, 13–21 (2009)
D.K. Mellinger, *Ishmael 1.0 User's Guide* (NOAA, Seattle, WA, 2001). Tech. Report OAR-PMEL-120
D.K. Mellinger, C.W. Clark, Methods for automatic detection of mysticete sounds. Mar. Freshwater Behav. Physiol. **29**, 163–181 (1997)
D.K. Mellinger, C.W. Clark, Recognizing transient low-frequency whale sounds by spectrogram correlation. J. Acoust. Soc. Am. **107**, 3518–3529 (2000)
D.K. Mellinger, B.A. Weisburn, S.G. Mitchell, C.W. Clark, Measuring regular whale call intervals with the summed autocorrelation. J. Acoust. Soc. Am. **95**, 2881 (1994)
D.K. Mellinger, K.M. Stafford, C.G. Fox, Seasonal occurrence of sperm whale (*Physeter macrocephalus*) sounds in the Gulf of Alaska, 1999–2001. Mar. Mamm. Sci. **20**, 48–62 (2004)
D.K. Mellinger, R.P. Morrissey, S.W. Martin, L. Thomas, T.A. Marques, J. Yosco, A method for detecting whistles, moans, and other frequency contours. J. Acoust. Soc. Am. **129**, 4055–4061 (2011)
J. Miksis-Olds, J.R. Buck, M.J. Noad, D.H. Cato, M.D. Stokes, Information theory analysis of Australian humpback whale song. J. Acoust. Soc. Am. **124**, 2385–2393 (2008)
P.J. Miller, P.L. Tyack, A small towed beamforming array to identify vocalizing resident killer whales (*Orcinus orca*) concurrent with focal behavioral observations. Deep-Sea Res. **45**, 1389–1405 (1998)
T. Mitchell, *Machine Learning* (McGraw-Hill, Boston, MA, 1997)
T.K. Moon, The expectation-maximization algorithm. IEEE Signal Proc. Mag. **13**, 47–60 (1996)
R.P. Morrissey, J. Ward, N. DiMarzio, S. Jarvis, D.J. Moretti, Passive acoustic detection and localization of sperm whales (*Physeter macrocephalus*) in the Tongue of the Ocean. Appl. Acoust. **67**, 1091–1105 (2006)
L.M. Munger, D.K. Mellinger, S.M. Wiggins, S.E. Moore, J.A. Hildebrand, Performance of spectrogram correlation in detecting right whale calls in long-term recordings from the Bering Sea. Can. Acoust. **33**, 25–34 (2005)
A.E. Newhall, Y.-T. Ying, J.F. Lynch, M.F. Baumgartner, G.G. Gawarkiewicz, Long distance passive localization of vocalizing sei whales using an acoustic normal mode approach. J. Acoust. Soc. Am. **131**, 1814–1825 (2012)
NIST. U.S. Natl. Inst. Standards Tech., 2010. http://www.itl.nist.gov/iad/mig/tools/. Accessed 4 Mar 2010
M.J. Noad, D.H. Cato, Swimming speeds of singing and non-singing humpback whales during migration. Mar. Mamm. Sci. **23**, 481–495 (2007)
E.-M. Nosal, L.N. Frazer, Delays between direct and reflected arrivals used to track a single sperm whale. Appl. Acoust. **87**, 1187–1201 (2006)
E.-M. Nosal, Methods for tracking multiple marine mammals with wide-baseline passive acoustic arrays. J. Acoust. Soc. Am. **134**, 2383–2392 (2013)
E.M. Oleson, S.M. Wiggins, J.A. Hildebrand, Temporal separation of blue whale call types on a southern California feeding ground. Anim. Behav. **74**, 881–894 (2007)
A.V. Oppenheim, R.W. Schafer, *Discrete-Time Signal Processing*, 3rd edn. (Prentice-Hall, Englewood Cliffs, NJ, 2009)
J.N. Oswald, S. Rankin, J. Barlow, M.O. Lammers, A tool for real-time acoustic species identification of delphinid whistles. J. Acoust. Soc. Am. **122**, 587–595 (2007)
J.D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Kruger, A.E. Lefohn, T.J. Purcell, A survey of general-purpose computation on graphics hardware. Comput. Graph. Forum **26**, 80–113 (2007)
J.R. Potter, D.K. Mellinger, C.W. Clark, Marine mammal call discrimination using artificial neural networks. J. Acoust. Soc. Am. **96**, 1255–1262 (1994)
S. Qian, D. Chen, Joint time-frequency analysis. IEEE Signal Proc. Mag. **16**(2), 52–67 (1999)

R Development Core Team, *R: A Language and Environment for Statistical Computing* (World Wide Web electronic publication, 2009). http://www.r-project.org/. Accessed 28 Jan 2010
L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**, 257–286 (1989)
L.R. Rabiner, B.-H. Juang, *Fundamentals of Speech Recognition* (Prentice Hall, Englewood Cliffs, NJ, 1993)
M.H. Rasmussen, L.A. Miller, W.W.L. Au, Source levels of clicks from free-ranging white beaked dolphins (*Lagenorhynchus albirostris* Gray 1846) recorded in Icelandic waters. J. Acoust. Soc. Am. **111**, 1122–1125 (2002)
B.P. Rideout, S.E. Dosso, D.E. Hannay, Underwater passive acoustic localization of Pacific walruses in the northeastern Chukchi Sea. J. Acoust. Soc. Am. **134**, 2534–2545 (2013)
D. Risch, C.W. Clark, P.J. Dugan, M. Popescu, U. Siebert, S.M. Van Parijs, Minke whale acoustic behavior and multi-year seasonal and diel vocalization patterns in Massachusetts Bay, USA. Mar. Ecol. Prog. Ser. **489**, 279–295 (2013)
M.A. Roch, M.S. Soldevilla, J.C. Burtenshaw, E.E. Henderson, J.A. Hildebrand, Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California. J. Acoust. Soc. Am. **121**, 1737–1748 (2007)
M.A. Roch, M.S. Soldevilla, R. Hoenigman, S.M. Wiggins, J.A. Hildebrand, Comparison of machine learning techniques for the classification of echolocation clicks from three species of odontocetes. Can. Acoust. **36**, 41–47 (2008)
M.A. Roch, H. Klinck, S. Baumann-Pickering, D.K. Mellinger, S. Qui, M.S. Soldevilla, J.A. Hildebrand, Classification of echolocation clicks from odontocetes in the Southern California Bight. J. Acoust. Soc. Am. **129**, 467–475 (2011a)
M.A. Roch, T.S. Brandes, B. Patel, Y. Barkley, S. Baumann-Pickering, M.S. Soldevilla, Automated extraction of odontocete whistle contours. J. Acoust. Soc. Am. **130**, 2212–2223 (2011b)
L.S. Sayigh, P.L. Tyack, R.S. Wells, Recording underwater sounds of free-ranging dolphins while underway in a small boat. Mar. Mamm. Sci. **9**, 209–213 (1993)
R.O. Schmidt, A new approach to geometry of range difference location. IEEE Trans. Aerosp. Electron. Syst. **8**, 821–835 (1972)
G. Sharma, J. Martin, MATLAB: a language for parallel computing. Int. J. Parallel Prog. **37**, 3–36 (2009)
A.D. Shapiro, P.L. Tyack, S. Seneff, Comparing call-based versus subunit-based methods for categorizing Norwegian killer whale, *Orcinus orca*, vocalizations. Anim. Behav. **81**, 377–386 (2011)
A. Širović, J.A. Hildebrand, S.M. Wiggins, Blue and fin whale call source levels and propagation range in the Southern Ocean. J. Acoust. Soc. Am. **122**, 1208–1215 (2007)
M.D. Skowronski, J.G. Harris, Acoustic detection and classification of microchiroptera using machine learning: lessons learned from automatic speech recognition. J. Acoust. Soc. Am. **119**, 1817–1833 (2006)
M.S. Soldevilla, E.E. Henderson, G.S. Campbell, S.M. Wiggins, J.A. Hildebrand, M.A. Roch, Classification of Risso's and Pacific white-sided dolphins using spectral properties of echolocation clicks. J. Acoust. Soc. Am. **124**, 609–624 (2008)
J.L. Spiesberger, K.M. Fristrup, Passive localization of calling animals and sensing of their acoustic environment using acoustic tomography. Am. Nat. **135**, 107–153 (1990)
J.L. Spiesberger, Finding the right cross-correlation peak for locating sounds in multipath environments with a fourth-moment function. J. Acoust. Soc. Am. **108**, 1349–1352 (2000)
J.L. Spiesberger, Hyperbolic location errors due to insufficient numbers of receivers. J. Acoust. Soc. Am. **109**, 3076–3079 (2001)
J.L. Spiesberger, M. Wahlberg, Probability density functions for hyperbolic and isodiachronic locations. J. Acoust. Soc. Am. **112**, 3046–3052 (2002)
K.M. Stafford, C.G. Fox, D.S. Clark, Long-range acoustic detection and localization of blue whale calls in the northeast Pacific Ocean. J. Acoust. Soc. Am. **104**, 3616–3625 (1998)
R. Suzuki, J.R. Buck, P.L. Tyack, Information entropy of humpback whale songs. J. Acoust. Soc. Am. **119**, 1849–1866 (2006)

J.A. Swets, *Signal Detection and Recognition by Human Observers: Contemporary Readings* (Wiley, New York, NY, 1964)
A. Tarantola, *Inverse Problem Theory* (Elsevier, Amsterdam, 1987)
J.R. Taylor, *An Introduction to Error Analysis (Chapter 3)*, 2nd edn. (University Science Press, Sausalito, CA, 1997)
G.K. Thiruvathukal, Cluster computing. Comput. Sci. Eng. **7**, 11–13 (2005)
A. Thode, G.L. D'Spain, W.A. Kuperman, Matched-field processing, geoacoustic inversions, and source signature recovery of blue whale vocalizations. J. Acoust. Soc. Am. **107**, 1286–1300 (2000)
A.M. Thode, D.K. Mellinger, S. Stienessen, A. Martinez, K. Mullin, Depth-dependent acoustic features of diving sperm whales (*Physeter macrocephalus*) in the Gulf of Mexico. J. Acoust. Soc. Am. **112**, 308–321 (2002)
A. Thode, Three-dimensional passive acoustic tracking of sperm whales (*Physeter macrocephalus*) in ray-refracting environments. J. Acoust. Soc. Am. **18**, 3575–3584 (2005)
A. Thode, J. Skinner, P. Scott, J. Roswell, J. Straley, K. Folkert, Tracking sperm whales with a towed acoustic vector sensor. J. Acoust. Soc. Am. **128**, 2681–2694 (2010)
C.O. Tiemann, M.B. Porter, L.N. Frazer, Localization of marine mammals near Hawaii using an acoustic propagation model. J. Acoust. Soc. Am. **115**, 2834–2843 (2004)
C.O. Tiemann, S.W. Martin, J.R. Mobley, Aerial and acoustic marine mammal detection and localization on Navy ranges. IEEE J. Ocean Eng. **31**, 107–119 (2006)
C.O. Tiemann, A.M. Thode, J. Straley, V. O'Connell, K. Folkert, Three-dimensional localization of sperm whales using a single hydrophone. J. Acoust. Soc. Am. **120**, 2355–2365 (2007)
A. Tolstoy, *Matched Field Processing* (World Scientific, Singapore, 1993)
W.A. Tyrrell, Design of acoustic systems, in *Marine Bioacoustics*, ed. by W.N. Tavolga (Pergamon, Oxford, 1964), pp. 65–86
I.R. Urazghildiiev, C.W. Clark, Acoustic detection of North Atlantic right whale contact calls using the generalized likelihood ratio test. J. Acoust. Soc. Am. **120**, 1956–1963 (2006)
I.R. Urazghildiiev, C.W. Clark, T.P. Krein, S.E. Parks, Detection and recognition of North Atlantic right whale contact calls in the presence of ambient noise. IEEE J. Ocean Eng. **34**, 358–368 (2009)
M. Wahlberg, B. Møhl, P.T. Madsen, Estimating source position accuracy of a large-aperture hydrophone array for bioacoustics. J. Acoust. Soc. Am. **109**, 397–406 (2001)
W.A. Watkins, W.E. Schevill, Four-hydrophone array for acoustic three-dimensional localization. Woods Hole Oceanogr. Inst., Tech. Report WHOI-71-60, 1971
W.A. Watkins, W.E. Schevill, Sound source location by arrival-times on a non-rigid three dimensional hydrophone array. Deep-Sea Res. **19**, 691–706 (1972)
W.A. Watkins, M.A. Daher, G.M. Reppucci, J.E. George, D.L. Martin, N.A. DiMarzio, D.P. Gannon, Seasonality and distribution of whale calls in the North Pacific. Oceanography **13**(1), 62–67 (2000)
P.R. White, M.L. Hadley, Introduction to particle filters for tracking applications in the passive acoustic monitoring of cetaceans. Can. Acoust. **36**, 146–152 (2008)
H. Whitehead, J. Gordon, Methods of obtaining data for assessing and modelling sperm whale populations which do not depend on catches. Rep. Intl. Whal. Commn. **8(Special Issue)**, 149–166 (1986)
S. Wiggins, M. McDonald, L.M. Munger, S. Moore, J.A. Hildebrand, Waveguide propagation allows range estimates for North Pacific right whales in the Bering Sea. Can. Acoust. **32**, 146–154 (2004)
S.M. Wiggins, M.A. McDonald, J.A. Hildebrand, Beaked whale and dolphin tracking using a multichannel autonomous acoustic recorder. J. Acoust. Soc. Am. **131**, 156–163 (2012)
S.J. Young, G. Evermann, M.J.F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P.C. Woodland, *The HTK Book, Version 3.4* (University of Surrey, Guildford, 2006)

W.M.X. Zimmer, M.P. Johnson, A. D'Amico, P. Tyack, Combining data from a multisensory tag and passive sonar to determine the diving behavior of a sperm whale (*Physeter macrocephalus*). IEEE J. Ocean Eng. **28**, 13–28 (2003)

W.M.X. Zimmer, *Passive Acoustic Monitoring of Cetaceans* (Cambridge University Press, Cambridge, 2011)

W.M.X. Zimmer, Range estimation of cetaceans with compact volumetric arrays. J. Acoust. Soc. Am. **134**, 2610–2618 (2013)