Recognizing transient low-frequency whale sounds by spectrogram correlation

David K. Mellinger[a)]
Cooperative Institute for Marine Resources Studies, Oregon State University, 2030 South Marine Science Drive, Newport, Oregon 97365, Monterey Bay Aquarium Research Institute, P.O. Box 628, Moss Landing, California 95039, and Bioacoustics Research Program, Cornell University, 159 Sapsucker Woods Road, Ithaca, New York 14850-1999

Christopher W. Clark
Bioacoustics Research Program, Cornell University, 159 Sapsucker Woods Road, Ithaca, New York 14850-1999

(Received 25 January 1999; accepted for publication 26 February 2000)

A method is described for the automatic recognition of transient animal sounds. Automatic recognition can be used in wild animal research, including studies of behavior, population, and impact of anthropogenic noise. The method described here, spectrogram correlation, is well-suited to recognition of animal sounds consisting of tones and frequency sweeps. For a sound type of interest, a two-dimensional synthetic kernel is constructed and cross-correlated with a spectrogram of a recording, producing a recognition function—the likelihood at each point in time that the sound type was present. A threshold is applied to this function to obtain discrete detection events, instants at which the sound type of interest was likely to be present. An extension of this method handles the temporal variation commonly present in animal sounds. Spectrogram correlation was compared to three other methods that have been used for automatic call recognition: matched filters, neural networks, and hidden Markov models. The test data set consisted of bowhead whale (*Balaena mysticetus*) end notes from songs recorded in Alaska in 1986 and 1988. The method had a success rate of about 97.5% on this problem, and the comparison indicated that it could be especially useful for detecting a call type when relatively few (5–200) instances of the call type are known. [S0001-4966(00)01706-9]

PACS numbers: 43.80.Lb, 43.80.Ka [WA]

INTRODUCTION

Automatic recognition of animal vocalizations is promising as a tool for investigating free-ranging animals in their natural habitats. It is also challenging as a problem in signal processing, since difficulty arises from the nonstationary nature of the signals involved; from the highly variable nature of animal sounds at the individual, intraspecific, and interspecific levels of analysis; and from the characteristics of the noise environments encountered in field recordings.

Primary applications of acoustical methods for studying marine animals include species range and distribution surveys (Clark and Mellinger, 1994; Clark and Fristrup, 1996; Moore *et al.*, 1998; Stafford *et al.*, 1998), behavior studies (Buck and Tyack, 1993; Frankel *et al.*, 1995), population measurement (Leaper *et al.*, 1992; Zeh *et al.*, 1993; Raftery and Zeh, 1998), and studies of the impact of noise on animals (Clark *et al.*, 1998). The more traditional methods of approaching these problems involve visual observation, but acoustic recognition methods have certain advantages over visual ones, especially when augmented with techniques for tracking animals in their native habitat (Watkins and Schevill, 1972; Clark *et al.*, 1986; Clark, 1989; Clark *et al.*, 1991, 1996; McDonald *et al.*, 1995). Some environments, including dense forests and oceans, are inaccessible or inhospitable to visual observation but more accommodating to acoustic methods. Visual observation is possible only during daylight hours, while acoustic methods can be used for species that are active anytime during the 24-hour day. Such species include ones active nocturnally (Griffin, 1964; Graber, 1968; Larkin, 1978; Terres, 1980) or throughout the 24-hour day, such as right whales (*Eubalaena* sp.), bowheads, and fin whales (*Balaenoptera physalus*) (Clark, 1983; Watkins *et al.*, 1987; Würsig and Clark, 1993). With automated recognition methods, extensive sound recordings can be made and processed relatively quickly, while visual observation is usually very labor-intensive and requires trained observers. Conditions may make visual observation difficult, as in the case of fog, or ice cover in the Arctic. For many species, especially ones in the marine environment that rely heavily on the acoustic modality for communication and information gathering, acoustic methods may offer more appropriate insights into the biology of the animals than visual methods.

Acoustic recognition can be done by persons with trained ears, but automated methods have certain advantages. They are unbiased, or rather their bias is constant rather than possibly changing from time to time and place to place. They can be used to process large amounts of data; this is quite important in field work, where thousands of hours of sound may need to be analyzed. In real-time processing situations, automated methods can operate on many sound channels at once, allowing simultaneous monitoring of several widely

[a)]Electronic mail: mellinger@pmel.noaa.gov

distributed different locations, or of sounds from the same source received at several different sensors. Sounds above or below the frequency range of human hearing can be processed, as can sounds that change too quickly or too slowly for humans to hear clearly. Automatic methods may also be relatively inexpensive for long-term monitoring.

A variety of techniques has been used for automatic recognition of animal calls. Spectrogram matched filtering, or cross-correlation of the spectrograms of a known sound and an unknown recording, has been used for classifying and comparing animal sounds (Clark *et al.*, 1987; Chabot, 1988). A variant of matched filtering uses synthetic waveforms (Stafford *et al.*, 1998) or synthetic spectrograms (Mellinger and Clark, 1993) instead of sounds edited out of recordings. Another approach is to measure a number of characteristics from the sound and use these in a statistical classifier (Fristrup, 1992; Pinkowski, 1994); this method has been successful at recognizing calls in a large database of marine mammal sounds (Fristrup and Watkins, 1994). The frequency contour of a vocalization can be tracked (Newman *et al.*, 1978; Clark, 1982; Goedeking, 1983; Buck and Tyack, 1993; Nyamsi *et al.*, 1994) for comparison to other contours. Neural networks, typically using spectrogram values as input, work well in many cases for detecting the sounds of a species of interest (Ramani and Patrick, 1992; Gaetz *et al.*, 1993; Moore *et al.*, 1991; Patrick *et al.*, 1994). Speech methods (Rabiner and Juang, 1993), in particular hidden Markov models, have been applied as well (Weisburn *et al.*, 1993; Sturtivant and Datta, 1997). Methods employing machine learning in combination with statistical feature extraction have been used successfully for monitoring frog sounds in Australia (Taylor, 1995; Taylor *et al.*, 1996), and a variety of *ad hoc* signal-processing systems has been used as well (Whitehead and Weilgart, 1990; Leaper *et al.*, 1992).

What characteristics of animal vocalizations make the automatic recognition problem tractable? Many vocalizations have evolved to be heard against background noise, and thus contain changing parameters that make them audible to the perceptual systems receiving them. One type of change is frequency modulation (FM); that is, a change in a signal's narrow-band, instantaneous frequency over time. Narrow-band FM signals, also known as frequency sweeps, are found in the vocalizations of many animals. Many acoustic signals of primates incorporate FM (Moody and Stebbins, 1989), as do those of birds, whose long-range advertisement songs ''almost invariably consist of frequency-modulated tones'' (Wiley and Richards, 1982). Most cetacean species, comprising the toothed or odontocete whales and the baleen or mysticete whales, produce a great variety of sounds, covering an extensive frequency band and including both FM signals and pulse sequences (Schevill, 1964; Thompson *et al.*, 1979; Herman and Tavolga, 1980; Clark, 1990).

Whales of the mysticete group, which includes only 11 species, show a wide range of inter- and intraspecific variability in their vocalizations. Within the more accessible coastal species, the sounds from bowhead, humpback (*Megaptera novaeangliae*), and right whales are the best documented, and all produce a wide variety of FM and pulsed sound structures (Schevill and Watkins, 1962; Payne and McVay, 1971; Payne and Payne, 1971; Clark, 1981, 1982; Payne *et al.*, 1983; Clark and Johnson, 1984; Chabot, 1984; Silber, 1986; Clark, 1991). In contrast, within the *Balaenoptera* genus that includes the blue (*Balaenoptera musculus*), fin, and minke (*B. acutorostrata*) whale species, signal documentation has been more limited since access has been difficult (Schevill *et al.*, 1964; Cummings and Thompson, 1971; Winn and Perkins, 1976; Edds, 1982; Watkins *et al.*, 1987; Thompson *et al.*, 1996; Rivers, 1997; Stafford *et al.*, 1999).

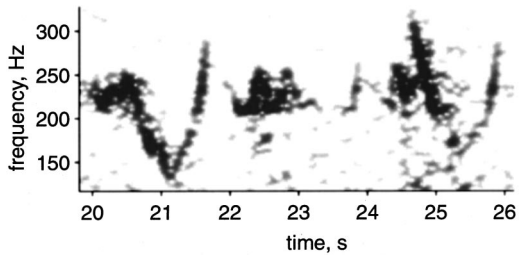

FIG. 1. Three successive bowhead song end notes from a song of one whale, recorded in Alaska in 1986. The first shows the frequency contour typical of end notes; the other two show typical variation. Spectrogram parameters for all figures: Sampling rate 2023 Hz, frame size 512 samples (zero-padded to an FFT size of 1024 samples), filter bandwidth 16 Hz, frame increment 32 ms, Hamming window.

The method described here for recognizing animal vocalizations focuses on finding frequency sweeps. The method was motivated by psychoacoustic work showing the salience of frequency sweeps in human hearing (Kay and Matthews, 1972), and by neurophysiological studies showing that there are neurons in the mammalian auditory system selectively responsive to frequency sweeps (Whitfield and Evans, 1965; Møller, 1977; Mendelson and Cynader, 1985; deCharms *et al.*, 1998). The method's image-processing technique is inspired by vision research that has identified neurons responsive to certain orientations of lines in the visual field (Hubel and Wiesel, 1962, 1977), analogous to certain orientations of frequency sweeps in the auditory field.

In this paper, we describe a specific call-recognition problem, present a solution in the form of a method for recognizing frequency sweeps, describe ways to tune parameters to make the method most effective, test the method on sets of sample data, compare results to those from other methods for animal sound recognition, and offer some improvements that raise performance of the method.

I. PROBLEM STATEMENT

The specific problem addressed here is the recognition of the end notes of bowhead whale songs. These notes are portions of a bowhead's song that occur one or more times in succession at the end of each song repetition, and are distinctly different from the preceding portions of the song. Songs by one individual typically last 60–70 s, with pauses of 5–15 s between songs (Würsig and Clark, 1993). Bowhead song end notes were chosen because they are relatively loud, and they typically occur several times per song (Clark, 1991). Figure 1 shows, in the leftmost of the three calls, a typical end note from the 1988 song for the Bering-Chukchi-Beaufort Sea stock of the bowhead whale, as well as two

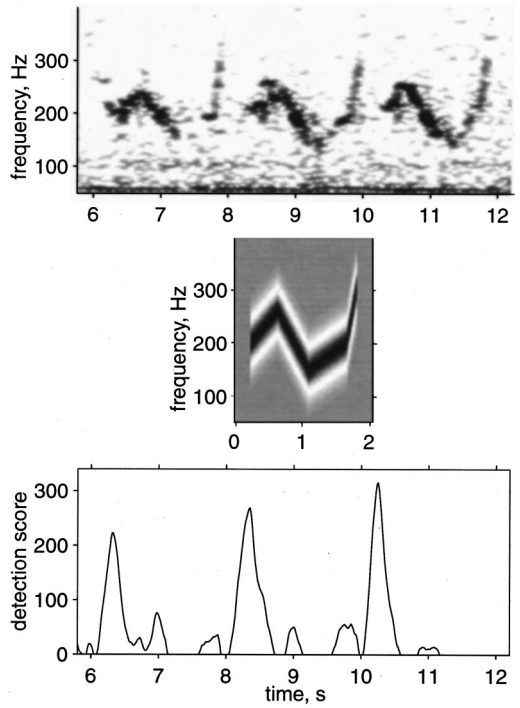

FIG. 2. Spectrogram correlation example. (a) Spectrogram of a novel recording containing three bowhead song end notes. (b) Spectrogram correlation kernel, with black representing positive areas, white representing negative areas, and gray representing the zero background. (c) The resulting recognition score function. The units of the recognition score function are arbitrary.

other end notes illustrating the amount of internote variation that can be present in sounds recorded from one whale. Recognizing sounds from bowhead whales provides a signal-processing challenge in that the sonic environment of the recordings contains many types of interfering noise, both stationary and transient.

II. METHOD

The method for recognizing animal vocalizations is described here in general terms, with specific application to the recognition of bowhead song end notes to make implementation of the fundamental technique concrete. The method operates on spectrograms computed from the time-series waveform of a sound (Oppenheim and Schafer, 1975; Altes, 1980). Examples of the sound type (the bowhead song end note, in this case) are used to construct a correlation kernel for the vocalization. To recognize the vocalization of interest in a novel recording, a spectrogram of the recording is made and cross-correlated with the kernel representing the signal of interest. The result is a recognition function—a time series of recognition values—that represents the closeness of the match between the kernel and the novel recording at each time increment of the spectrogram. Larger values in the time-varying recognition function represent higher likelihood that a bowhead song end note was present in the novel recording. An example of the process is shown in Fig. 2.

To use this method, a correlation kernel is made for the vocalization type of interest, as follows. A set of training examples of the vocalization of interest is obtained. Spectrogram time/frequency parameters are chosen that make the vocalization's frequency contours clearly visible in the spectrogram. Bradbury and Vehrencamp (1998) further discuss choice of spectrogram parameters for animal vocalizations. Our experience has been that parameters that reveal frequency contours for visual inspection also work well for spectrogram correlation. Next, the vocalization is divided into *sections* in which the FM rate is relatively constant. The kernel is made up of a series of *segments*, where each segment corresponds to one section of the vocalization. For instance, the first bowhead song end note of Fig. 1 has four sections: a short initial FM upsweep, a downsweep, an upsweep, and another, steeper, upsweep. The corresponding kernel made for this song end note therefore would have four segments.

TABLE I. Measured mean (±s.d.) characteristics of the bowhead song end notes from 1986; values were measured in a spectrogram computed with the same parameters as Fig. 1. The bandwidth value is an approximate estimation (see the text). The mean values were used for constructing the recognition kernel *k* containing four contiguous segments: an initial frequency upsweep, a downsweep, and two successive upsweeps. Since the frequency sweeps in this case were contiguous throughout the vocalization, the end frequency of each section is the same as the start frequency of the next section.

Section no.	Duration, s	Start freq., Hz	End freq., Hz	Bandwidth, Hz
1	0.395±0.079	197.6±9.3	257.2±15.1	50
2	0.458±0.110	257.2±15.1	146.4±12.2	50
3	0.557±0.139	146.4±12.2	196.7±15.5	50
4	0.143±0.105	196.7±15.5	295.7±19.4	50

To construct a kernel for the vocalizations in the training set, time and frequency characteristics of the sections that make up each sample vocalization are measured. The start time of the vocalization is set to zero. For each section in the vocalization, the times and frequencies of the section's endpoints are measured. This measurement may be done using widely available computerized spectrogram-display tools, such as CANARY (Charif *et al.*, 1995), OSPREY (Mellinger, 1995), or SIGNAL (Engineering Design, 1997). Also, the bandwidth of a each section is measured—the difference between the upper and lower edges of the section visible in the spectrogram at one instant. The measurements for each endpoint are averaged for all sample vocalizations to obtain characteristics of an ''average call'' that is used to make up the kernel.

For recognizing bowhead song end notes, songs from two whales recorded 5 days apart were used, and five end notes from each of four total songs were selected, for a total of 20 training notes. The endpoints and bandwidths of the four sections in each note were measured, then averaged. These averaged values are shown in Table I.

The kernel is made up of several segments, one per FM section in the target vocalization type. Each kernel segment is a two-dimensional array of values in the time–frequency space of spectrograms; it contains positive- and negative-valued regions. The kernel for the bowhead song end notes is shown in the center of Fig. 2. Its positive region extends between the averaged start and end frequencies of the measured FM sections, has a duration equal to the average duration of the sections, and has a frequency spread approxi-

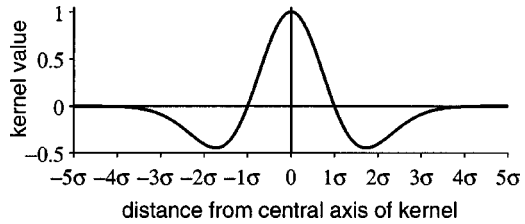

FIG. 3. Values in a vertical cross section of the kernel segment of Fig. 2. The horizontal axis is the distance from the axis of the FM sweep in the kernel, in units of the spread factor σ. The vertical axis is the kernel value, displayed here as a curve instead of gray-scale values as in Fig. 2.

mately equal to the bandwidth of the measured sections. Each segment also has two negative regions, at higher and lower frequency bands above and below the central positive region, respectively. These negative regions aid in the rejection of noise sounds. These are shown in the kernel of Fig. 2 as lighter areas. The shape of the positive and negative areas in the kernel is defined as a hat function, the second derivative of a Gaussian distribution as shown in Fig. 3. The kernel value k at a given time and frequency point (t,f) is specified by

$$x = f - \left(f_0 + \frac{t}{d}(f_1 - f_0) \right), \tag{1}$$

$$k(t,f) = \left(1 - \frac{x^2}{\sigma^2} \right) \exp\left(-\frac{x^2}{2\sigma^2} \right), \tag{2}$$

where x = distance of this (t,f) point from the central axis of the segment at time t, f_0 = start frequency of the segment, f_1 = end frequency of the segment, d = duration of the segment, and σ = instantaneous bandwidth of the segment at time t.

The positive and negative values of a segment sum to zero. This is important for rejection of noise and of interfering sounds. When a spectrogram of uniform background noise is made and cross-correlated with such a zero-sum kernel, the result is zero, since the noise aligns with equal regions of positive and negative values of the kernel. Similarly, if a frequency sweep of a different FM sweep rate (df/dt) is cross-correlated with this kernel, the sweep will intersect both positive and negative regions of the kernel, producing a summed output response of zero. Other types of noise such as wideband clicks and pops, commonly produced by both cracking ice and recording glitches, also intersect positive and negative regions and produces a summed output response of zero. As the next section shows, this characteristic has proven effective at rejecting many types of interfering noise—a useful feature in the acoustically cluttered arctic environment of the bowhead call-recognition example. A further feature is that many noisy sounds can co-occur with bowhead song end notes and not affect the correlation output, because the correlation operation is linear and the interfering sounds simply add zero to its output. Of course, if an interfering noise has frequency sweeps that closely match the sound of interest (the bowhead song end note, in our example), then the correlation value will be undesirably large.

The value of the bandwidth parameter σ may be varied to accommodate a greater or lesser amount of variation in the instantaneous bandwidth of frequency sweeps. Greater kernel bandwidth results in a less discriminatory recognition function that includes more types of frequency sweeps. That is, a kernel segment constructed with a large σ value, when cross-correlated with a wide variety of input frequency sweeps, will produce relatively high recognition values, while small bandwidths in the kernel segment will produce equivalent recognition values only for frequency sweeps that are more exactly aligned with the kernel axis. As a rule of thumb, σ values of 0.7 ± 0.2 times the instantaneous bandwidth of the actual frequency sweep have been found to work well. Since small variations ($\pm 10\%$) in σ were found to have little effect on recognizer performance, exact measurement and averaging of the instantaneous bandwidth of many frequency sweeps was not necessary. However, different choices of spectrogram parameters, especially fast Fourier transform (FFT) size and window function, will lead to different instantaneous bandwidths of frequency sweeps, and will require different values of σ.

To produce the correlation kernel, several segments are concatenated. Given the kernel $k(t,f)$ and the spectrogram, the output of the cross-correlation process for each frame of the spectrogram is given by

$$\alpha(t) = \sum_{t_0} \sum_{f} k(t_0,f) S(t - t_0, f), \tag{3}$$

where $S(t,f)$ is the spectrogram, and the limits of summation are specified by the size of k in time and frequency. Note that the correlation is in the time dimension only, not in frequency as would be used for two-dimensional cross-correlation. This one-dimensional correlation is significantly faster than a two-dimensional one, and works well provided that the vocalizations to be recognized are not frequency shifted (see Clark *et al.,* 1987). This is true of bowhead whale vocalizations, and of many other species' sounds as well.

The result of the cross-correlation operations is a time series with a ''recognition score'' $\alpha(t)$ for each point in time, i.e., a value representing the closeness of match between the bowhead song end note and the kernel. The bottom of Fig. 2 shows an example of the recognition score function. This function, $\alpha(t)$, is not strictly a probability, as it does not vary between only 0 and 1, but its value is near 0 when a bowhead vocalization is absent and increases in value when a vocalization is present. A zero minimum is applied to the recognition score, that is, negative values are changed to zero. Because spectrogram levels are not normalized before the cross-correlation is computed, the maximum value of $\alpha(t)$ is arbitrarily large. Also, because of the definition given by Eq. (3), the time in the recognition function when the peak occurs is at the very beginning of each song end note.

The recognition score function $\alpha(t)$ may be turned into a sequence of discrete *detection events*—occasions when the sound of interest is determined to be present—by setting a threshold and registering an event each time the score goes above the threshold.

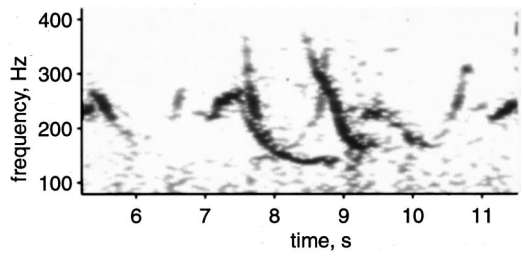

FIG. 4. Bearded seal sounds with some similarities to bowhead song end notes. The two prominent curving frequency downsweeps from 7.5–9 s and from 8.5–9.5 s are bearded seal sounds. The first of these has a frequency contour similar to sections of the end notes of Fig. 1.

III. RESULTS

A. Qualitative analysis

As part of an acoustic census program, bowhead whales were recorded in Spring 1986 and 1988 off Pt. Barrow, Alaska during their seasonal migration (Clark and Ellison, 1989; Clark *et al.*, 1996; Clark and Ellison, in press). Sounds were received by hydrophones placed along the edge of the shorefast ice in water 15–20 m deep, transmitted by FM radio to a field recording hut, and recorded with a Teac R-61D cassette data recorder. The recordings were later brought into the laboratory and digitized by a Teac RD-135T digital tape recorder for computer analysis. The recording systems' frequency responses were flat in the 14–5000 Hz frequency band.

The 1986 recordings contain bowhead songs that include the end notes of the type shown at the left in Fig. 1. The recordings also contain a variety of types of noise in the same frequency range as the bowhead song end notes: hydrophone cable flutter, wind and wave noise, ice noise (including squeals, creaks, and pops), and vocalizations of beluga whales (*Delphinapterus leucas*), bearded seals (*Erignathus barbatus*), and other bowhead whales.

A kernel was designed for the bowhead song end notes using the methods described above, and the spectrogram-correlation operation of Eq. (3) was computed. In preliminary tests, the spectrogram correlation method worked well at recognizing bowhead song end notes, producing high recognition scores when a note was present and low values at other times. Sometimes notes were missing one or more FM sections, in which case the recognition score was lower but still high relative to background noise.

The rating of a recognition score as either ''low'' or ''high'' depends on the size of the kernel, the gain of the sound acquisition system, and other factors. They are essentially arbitrary values that can be easily determined by inspection of the detector output. In this case a low score is one below 50, and a high score one above 100. Recognition scores, as expected, were low for white and colored background noise. The scores were also low for most types of interfering sounds, including hydrophone cable flutter, ice sounds, and most other marine mammal sounds. A few bearded seal sounds did result in high values ($\approx$150) of the spectrogram correlation in the correlation output. These sounds, two of which are shown in Fig. 4, had frequency contours similar to those of bowhead song end notes. This result is not altogether surprising, as human listeners often confused this type of seal sound with bowhead sounds.

The handling of variation in frequency contours is discussed more extensively and quantitatively in Sec. IV, below. Qualitatively, the method gave reliable results for bowhead song end notes that contained some variation in the frequency contour, but not a large amount. The amount of frequency variation that could be accommodated depended primarily on the bandwidth σ of the kernel segments as discussed above. As expected, variation in the duration of an end-note section, or in the timing of different sections, caused some problems for the method. A small amount of timing variation was accommodated by the nature of the up- and downsweeps of the kernel. For example, even when one of the FM sections in a bowhead song end note was slightly delayed, it still intersected part of the corresponding positive segment of the kernel.

B. Quantitative comparison

To test the spectrogram-correlation method, its performance was compared against that of a matched filter and a hidden Markov model (HMM) using a set of sample data. The sample data set for the comparison was selected by examining spectrograms of the Pt. Barrow recordings to determine when bowhead song end notes were present. Recordings were divided into short samples of a few seconds duration, with each sample containing a bowhead song end note. A total of 114 samples was selected this way.

Matched filtering is the optimum method for detecting a known signal in white Gaussian noise (van Trees, 1968). When detecting animal sounds, however, the signal is not known precisely since it varies from one occurrence to the next, and the background noise rarely if ever has a flat spectrum, so a matched filter is not necessarily the optimal solution. To compute the matched filter kernel, the time-series waveforms of several high-quality (good signal-to-noise ratio) notes were selected. Each time series was normalized to have zero mean and unit energy, and all time series were placed in a matrix. A sample covariance matrix was calculated, and the eigenvector corresponding to the maximum eigenvalue was computed. This calculation corresponds to a principal components projection of the notes (Jain and Dubes, 1988). The resulting time-series kernel was then cross-correlated with each of the test sound samples to produce output functions. As with the spectrogram kernel-correlation method, the maximum correlation value for a given sound sample was used as the recognition score.

HMM techniques have been used widely in speech recognition (Rabiner and Juang, 1986; Lee *et al.*, 1990; Rabiner and Juang, 1993) and it was thought that an HMM would perform well for the task of recognizing the complexity of bowhead song notes. The HMM used has been described elsewhere (Weisburn *et al.*, 1993); briefly, it was trained by picking peaks in several example calls and using their frequencies to trigger state transitions in the model.

The spectrogram correlation, matched filter, and HMM methods each produced a recognition score for each of the 114 sounds in the sample set. In order to determine if the score would be considered a detection event, it was neces-

TABLE II. Comparison of three recognition methods for a set of 114 bowhead song end notes recorded in 1986. The detection thresholds were set as described in the text.

Method	Correct detections	Error rate
Matched filter	96	15.8%
Hidden Markov model	111	2.6%
Spectrogram correlation	113	0.9%

sary to choose a detection threshold for each of the three methods, such that scores above the threshold would be considered a detection event.

The choice of a detection threshold depends on the application. Choosing a low detection threshold causes few localizations to be missed, but can also cause more background noise sounds to be falsely ''detected.'' Choosing a high threshold eliminates some or all of the false detections, but can also cause some actual vocalizations to be missed. If the aim is to scan a large body of recordings looking for a vocalization that may possibly be present, then choosing a low detection threshold makes sense, so that the rare sound is not missed. Any sounds detected can be re-examined later to see if they are the vocalizations of interest. Alternatively, automatic recognition can be used in an acoustic census to count the number of calls present in a set of recordings—typically, a large number of calls. In this case, missing some fraction of the total calls is not a severe drawback, as long as the fraction missed is known. Counting background noise sounds as calls, however, can affect the census result, so a high detection threshold is more appropriate.

For the problem of recognizing bowhead song end notes, it was decided to use a relatively low threshold, so that 10% of noise samples would be detected as calls. Thresholds for each of the methods were chosen by synthesizing 114 white-noise sound samples (generated using a normally distributed pseudorandom number generator), using them as input to the algorithms, and setting the threshold for each method so that 10% of the white-noise sounds were detected as calls and the remaining 90% rejected.

Using these detection thresholds, the 114 bowhead calls were processed by each of the recognition methods. The numbers of detection events for each method are shown in Table II. These results show that the matched filter worked poorly at recognizing bowhead song end notes, missing 18 of them. In contrast, the HMM worked fairly well, missing only 3 of the 114 notes, and the spectrogram correlator worked best of all, missing only 1.

C. Quantitative comparison to a neural network

Performance of the spectrogram-correlation method was compared to that of a neural network. For this comparison, a more extensive set of bowhead song end note recordings, made at Pt. Barrow in 1988, was used. Recording methods and equipment were similar to those in 1986, and the origins of noise were the same. This data set contains end notes from seven whales. Songs were identified as coming from an individual by acoustic tracking and by the timing of repeated songs. Spectrograms of the recordings were again examined to determine when bowhead song end notes were present, and recordings were edited to produce a set of 588 samples, each one 3.5 s in duration. Test samples that did not contain calls were also obtained; this was accomplished by editing out 3.5-s samples from Pt. Barrow recordings which did not contain bowhead sounds. A total of 888 noise samples containing ambient background noise, bearded seal sounds, ice noise, hydrophone flutter, and other sounds was obtained by this procedure.

Neural networks (Hebb, 1949; Rumelhart *et al.*, 1987) are often used in pattern-recognition systems where flexibility in the recognition of highly variable signals is required (Lippman, 1989; Ghosh *et al.*, 1992). The neural network architecture used for this comparison has been described elsewhere (Potter *et al.*, 1994). Briefly, it was a nonlinear network trained by backpropagation with momentum; the learning rate and tolerance values were decreased throughout the training period. The input layer was a spectrogram computed with a frequency range of 63–700 Hz, a filter bandwidth of 63.5 Hz, a duration of 2.7 s, and a time resolution of 0.128 s. A few selected pixels were deleted from the 11×21 spectrogram grid for a total of 192 input elements. The network's hidden layer had four units, and the output layer had a single unit. This output unit's value was compared to a threshold to determine whether or not a detection event had occurred. For training the network, each bowhead sound sample was time-aligned; noise samples were assigned a random time alignment. Approximately half of the data set was used for training, with the other half used as test data.

For an additional comparison, a matched filter was used to detect the bowhead song end notes as well. The matched-filter kernel for the 1988 end notes was constructed by the same method described above for the 1986 data.

The spectrogram correlation kernel was designed using the techniques described previously. Bowhead song end notes from 1988 had only three FM sections that were consistently present: a downsweep, followed by an upsweep, followed by a downsweep. Time and frequency values were measured for 48 calls from the training set. Time and fre-

TABLE III. Characteristics of the kernel used for recognizing bowhead song end notes from 1988. The kernel contains three segments, representing an initial FM downsweep, an upsweep, and a final downsweep. The FM sections in this case were not contiguous.

Section no.	Start time, s	End time, s	Start freq., Hz	End freq., Hz	Bandwidth, Hz
1	0	0.580	439.5	156.7	50
2	0.862	1.297	178.1	616.3	50
3	1.693	2.106	637.7	260.0	50

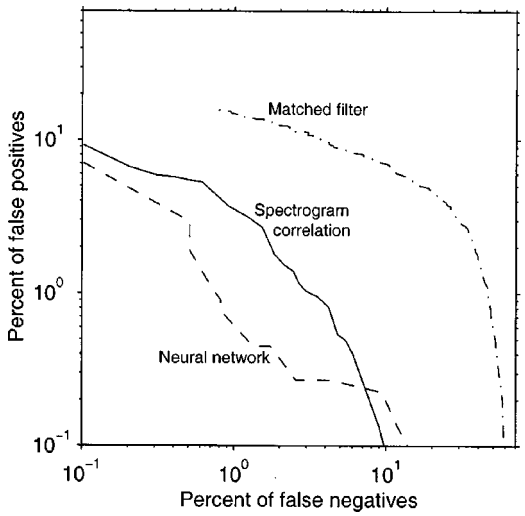

FIG. 5. Recognition error rates for the methods of spectrogram correlation, matched filter, and neural networks. Since error rates are plotted, curves lower and farther left represent better recognition performance.

quency averages are shown in Table III, along with the bandwidth used.

For this comparison many threshold values were used, instead of a single value as was done with the 1986 data. With a given setting of the threshold, two types of errors are possible. If a bowhead song end note is present in a sound and a recognition algorithm does not produce a recognition score above the threshold (and hence does not register a detection event), a ''false negative'' error occurs. If an end note is not present but the maximum recognition score is above threshold, a ''false positive'' error occurs. The false-positive and false-negative error rates are inversely related: if the threshold is raised, the number of false positive errors declines, but the number of false negative errors rises, and conversely for lowering the threshold. By setting the threshold to a series of increasing values, a succession of false-positive and false-negative error rates are obtained. The false-negative error rate may be plotted against the false-positive rate. Results of running the three recognizers—the spectrogram correlator, the matched filter, and the neural network—are shown in Fig. 5. In this figure, smaller error rates are represented by points toward the left and bottom of the plot.

As the figure shows, the matched-filter method worked only marginally well. The combined error rate, defined as the minimum of the sum of the false-positive and false-negative error rates, was 13.4% (9.0% false positive+4.5% false negative), a value too high for the matched filter to be of practical value.

The spectrogram-correlation method worked significantly better than the matched filter, perhaps well enough to make the method practical for some applications. The combined error rate was 3.6% (1.8% false positive+1.8% false negative). At this rate, the method could be used to screen long recordings for calls of interest.

The neural network had better performance, achieving a combined error rate of 1.6% (0.8% false positive+0.8% false negative). This error rate makes it the most accurate of the three methods discussed here for recognizing bowhead song end notes. As discussed below, however, there are occasions when spectrogram correlation is the method of choice.

D. Computational issues

The most computationally intensive automatic call-recognition task is detecting calls of interest in a large body of sound. For instance, the recognition system discussed above could be employed to detect bowhead whale song end notes in several months of continuous recordings spanning one field season. High speed is important for such intensive applications.

How does spectrogram correlation compare in speed to other methods? As an example, in operation on the set of bowhead end notes from 1988, spectrogram correlation ran approximately 2.6 times as fast as the matched filter, and about the same speed as the neural network. (These times are for operation only, not including training.) The higher speed of spectrogram correlation and neural network probably occurred because their slowest step is calculating FFTs in making a spectrogram. Further steps involving dot products of two-dimensional images (the spectrogram correlation kernel or the input layer of the neural network) actually involved relatively few spectrogram cells in this case and were faster than the calculation of the spectrogram. Matched filtering also requires an FFT (and inverse FFT), but on a much longer sequence of sound samples, which have a much higher sampling rate than spectrogram frames. Spectrogram correlation and neural networks will slow down as the size of the kernel increases, but in use on many different projects to date, its speed has always been comparable to or faster than matched filtering.

IV. TIME VARIATION

A. Problem description

One way in which the spectrogram-correlation method fails to work effectively is in handling timing variations within animal calls. Figure 1 shows bowhead song end notes that illustrate the large range of variation present in the data set. Some better means of accommodating the variation between calls might improve the performance of the automatic call recognizer. A similar problem exists in speech-recognition research, and several methods have been used to address it, including HMMs and dynamic time warping (Rabiner and Juang, 1993). The method used here (Mellinger, 1993) is a variant of dynamic time warping; it is described for a three-part bowhead song end note, but may be extended in an obvious way to sounds with any number of parts.

A call type is modeled, as above, as a sequence of separate sections, but now each section is separated from the next by an amount of time described by a probability distribution. For instance, the bowhead song end notes from 1988 can be modeled as three sections, as described previously. We now relax the assumption that the three sections occur in immediate succession, and instead describe them as separate FM sections occurring at different points in time. The amount of time between section i of the call and section $i+1$ is speci-

fied by $\lambda_i(t)$, a continuous distribution representing the probability that the two sections occur separated by time t.

To make a recognizer for a call type modeled this way, first recognizers for each separate section are constructed. Any type of recognizer works for processing these separate sections, provided that its output is a recognition score as a function of time. Here, a spectrogram correlator is assumed; constructing a spectrogram correlator amounts to making its kernel by the method outlined previously. Processing the training samples with these separate recognizers produces several recognition functions for each sample, one per section. Next, the time-delay probability distributions λ_i are estimated by measuring, for each training sample, times between the peaks of the recognition functions for the segments in the sound sample. The result is a set of time delays, one per training sample, that is used to estimate the time-delay probability functions. Typically, these time-delay probability estimates have a shape similar to a bell curve, and the distributions are therefore modeled as Gaussian.

The single-section recognizers and the time-delay probability functions make up the specification of the recognizer for the whole call type. After the separate section recognizers are run, the output functions, one per section, and the time-delay probability distributions are combined in a way described below to produce a single output recognition function.

To make a recognizer for the 1988 bowhead song end notes, three recognition kernels were made, one for each of the three sections. The kernels were constructed in the usual way, using measured time-frequency points from the training data of 48 end notes and zero-sum excitatory/inhibitory functions. Each input sound was correlated with the kernels as in Eq. (3), to produce three recognition functions $\alpha_1(t)$ through $\alpha_3(t)$. Figure 6 shows an example of the three single-segment kernels and associated recognition scores.

Note that in $\alpha_1(t)$, the function computed by spectrogram correlation with the first kernel, the peak occurs early in the end note, at about 0.85 s, where the first FM section of the bowhead sound begins. The peak in $\alpha_2(t)$ occurs later, at about 1.75 s, and the peak in $\alpha_3(t)$ occurs still later, at about 2.5 s. The differences between the times of these successive peaks are examples of the measurements used to estimate the distributions $\lambda_1(t)$ and $\lambda_2(t)$. These distributions were estimated from measurements of the trial set of 48 end notes. A histogram of the measured times between the initial upsweep and the downsweep—a sampling of the probability distribution $\lambda_1(t)$—is shown in Fig. 7. The assumption was made that the distribution was Gaussian and could be described by its mean and variance. This assumption was not required—the method will work with any distribution—but if the type of distribution is known, far fewer samples are needed to estimate it.

The recognition functions $\alpha_i(t)$ are combined to make a single master recognition score α

$$\alpha = \max_{t_1,t_2,t_3} (\alpha_1(t_1)+\lambda_1(t_2-t_1)\alpha_2(t_2) + \lambda_2(t_3-t_2)\alpha_3(t_3)). \quad (4)$$

Every possible time of occurrence t_i of each section of the bowhead song end note is used in this maximization. In it, the products of the recognition score for that section at that time $[\alpha_i(t)]$ and the probability of the associated time delay $[\lambda_i(t)]$ are computed. These products are summed, and the maximum of all such sums of products is taken as the overall recognition score.

The combination function α has the characteristic that whenever the separate recognition functions $\alpha_i(t)$ have peaks separated by the correct time delays, then α has a large value. If the recognition functions $\alpha_i(t)$ do not have large values, or if the delays between their large values differ from the high values in the distributions $\lambda_i(t)$, then α's value is small. Equation (4) defines α for a three-section call type, but it can easily be extended to sounds with any number of sections. Essentially, α is large when sections of the right

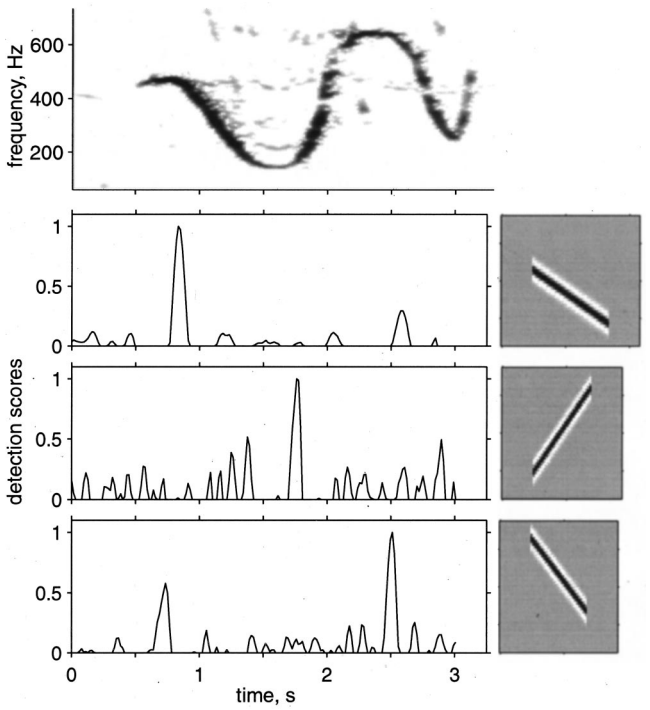

FIG. 6. Bowhead song end note from 1988, with (a) spectrogram and spectrogram correlator recognition functions and kernels for (b) the initial downsweep, (c) the middle upsweep, and (d) the final downsweep. Note the progressive delays in timing between the peaks in (b) and (c), and the peaks in (c) and (d). Recognition functions are scaled to have a maximum value of 1. In the images, the axes of the four kernels are not precisely parallel to the axes of the frequency sweeps in the bowhead song end note because the parameters used to construct this kernel were averaged over many bowhead song end notes.

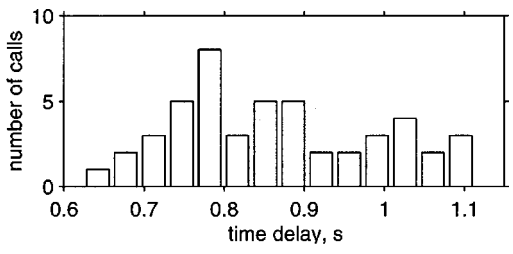

FIG. 7. Histogram of time delays between the first two sections of 1988 bowhead song end notes, as measured from the 48 end notes used for training.

type occur [the $\alpha_i(t)$ are large], and they occur separated by the right time delays [the $\lambda_i(t)$ are large].

As before, a threshold may be applied to this recognition score to determine detection events and then the associated false-positive and false-negative error rates.

B. Computational issues

Equation (4) represents a three-dimensional optimization problem, one that can be quite computationally expensive. Furthermore, the cost increases exponentially with the number of FM sections in the sound type, so that sound types only a bit more complex are very computationally expensive. This cost could be a problem, since one of the uses of automatic recognition methods is real-time monitoring of animal sounds. Fortunately, the three-dimensional calculation of Eq. (4) can be simplified to two two-dimensional calculations. In general, an n-dimensional calculation similar to Eq. (4) is reduced to $n-1$ two-dimensional calculations, greatly easing the computational burden.

To do this, note that the variable t_1 in Eq. (4) is used in only the first two terms. Since these terms are summed in computing the maximum, they could be maximized separately from the rest of the equation *if* they were independent of it. They are not independent, but the maximization may still be done separately using the Viterbi algorithm (Rabiner and Juang, 1993), a form of dynamic programming. Define

$$f_2(t_2)=\max_{t_1}(\alpha_1(t_1)+\lambda_1(t_2-t_1)\alpha_2(t_2)). \tag{5}$$

Then, from Eq. (4)

$$\alpha=\max_{t_2,t_3}(f_2(t_2)+\lambda_2(t_3-t_2)\alpha_3(t_3)). \tag{6}$$

Similarly,

$$f_3(t_3)=\max_{t_2}(f_2(t_2)+\lambda_2(t_3-t_2)\alpha_3(t_3)), \tag{7}$$

and

$$\alpha=\max_{t_3}(f_3(t_3)). \tag{8}$$

Computationally, the function $f_2(t_2)$ is simple to calculate: for each t_2, the value of t_1 that gives the maximum f_2 is calculated. Similarly, once $f_2(t_2)$ is known, $f_3(t_3)$ may be calculated separately for each value of t_3. Last, α can be calculated as a one-dimensional optimization problem.

Two steps of this procedure are two-dimensional computations: calculating $f_2(t_2)$ requires a one-dimensional maximization (over t_1) for each value of t_2, so for all values of t_2 this is a two-dimensional optimization. Calculating $f_3(t_3)$ is similarly two-dimensional, and the final maximization of α is only a one-dimensional calculation. Thus, this optimization method reduces the problem in complexity by a factor of $n/2$, where n is the number of time steps (frames) in the spectrogram. More generally, the computational complexity of a sound type with k FM sections is reduced by $n^{k-2}/(k-1)$.

With the above optimization, this method for handling time variation takes only a fraction of the time required for calculating a spectrogram. Fundamentally, the method operates on far fewer data points than are used in a spectrogram: it operates on only one data point per spectrogram frame, while the spectrogram calculation computes a many-point FFT for each spectrogram frame.

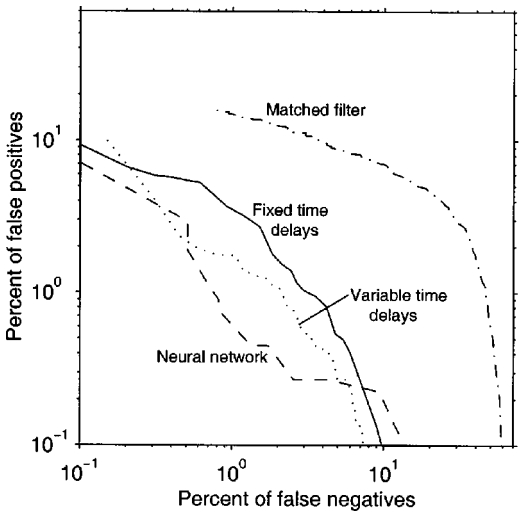

FIG. 8. Comparison of fixed-time method [Eq. (3), the same curve as the ''spectrogram correlation'' curve in Fig. 5] and variable time-delay method [Eq. (4)]. Curves lower and farther to the left represent lower error rates. Curves for the neural network and matched filter are plotted for reference.

C. Comparison

Does this new method, allowing timing variation among sections of a call, result in a recognizer that is any better at recognizing bowhead song end notes than the fixed-timing version of Eq. (3)?

Equations (5)–(8) were used on the same data set of end notes described in Sec. III. The false-positive versus false-negative curves for the time-variation method are shown in Fig. 8. Again, raising the threshold reduces the false-positive error rate, and lowering the threshold reduces the false-negative rate.

As the figure shows, the new method reduces the error rate moderately over most of the range of measurement. The combined error rate was 2.5% (1.9% false positive, 0.6% false negative). This represents an improvement of about one-third from the spectrogram-correlation error rate of 3.6%. The amount of improvement expected for other sound types depends on the amount of sound-to-sound variation, with sounds that vary more in timing between their constituent frequency sweeps likely to improve the most with this time-varying method.

D. Other combination functions

Equation (4) combines the terms $\lambda_i(t)\alpha_i(t)$ by adding them, but addition is not the only combination function possible. Any function that is monotonic and nondecreasing on positive numbers will work. Monotonicity is necessary for the optimization of Eqs. (5)–(8).

Two other combination functions were tried: multiplication and weighted addition, represented, respectively, by

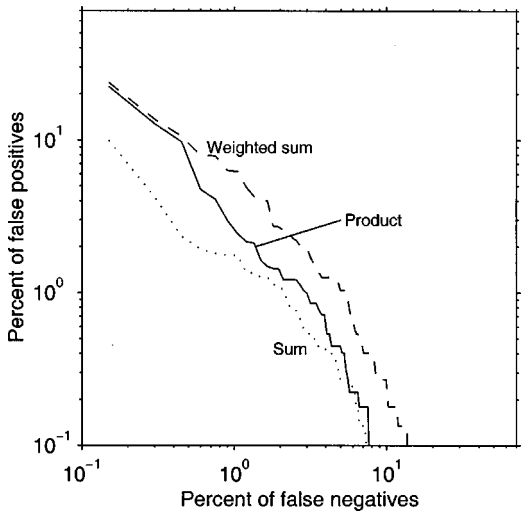

FIG. 9. Comparison of several combination rules for the time-varying recognition method. Results are shown for the sum method [Eq. (4), the same curve as the ''variable time-delay'' curve of Fig. 8], the product method [Eq. (9)], and the weighted-sum method [Eq. (10)].

$$\alpha = \max_{t_1,t_2,t_3} (\alpha_1(t_1) \cdot \lambda_1(t_2 - t_1)\alpha_2(t_2) \cdot \lambda_2(t_3 - t_2)\alpha_3(t_3)), \tag{9}$$

and

$$\alpha = \max_{t_1,t_2,t_3} (\alpha_1(t_1)w_1 + \lambda_1(t_2 - t_1)\alpha_2(t_2)w_2 + \lambda_2(t_3 - t_2)\alpha_3(t_3)w_3), \tag{10}$$

where each w_i is a weighting value chosen for each time delay.

The weighted-sum method was devised in response to the perceived importance of different sections of bowhead song end notes, which varied from section to section. The initial FM downsweep and the FM upsweep were noticed to be nearly always present and relatively loud. The third final FM downsweep was sometimes weak or absent. Accordingly, weights of $w_1 = 1.0$, $w_2 = 1.0$, and $w_3 = 0.5$ were somewhat arbitrarily assigned to the three sections, respectively. These equations were then applied to the data set of bowhead song end notes.

Results are shown in Fig. 9, where the sum [Eq. (4)], product [Eq. (9)], and weighted-sum [Eq. (10)] combination methods are compared. The product and weighted-sum methods were worse than the sum method over almost all of their range. The best combined error rates for the product method was 3.1% (1.6% false positive+1.5% false negative). The weighted-sum method was worse than the fixed-offset method of Sec. II, with a combined error rate of 4.5% (2.7% false positive, 1.8% false negative).

V. DISCUSSION

The spectrogram-correlation method presented here is effective at finding a large percentage of the bowhead sounds in a test database, much better than a simple matched filter. A neural-network method performs still better; however, there are several factors to consider in using one of these methods. Here, we outline the advantages and disadvantages of the recognition methods presented [using the sum version of Eq. (4)] and the other methods used for the comparison in Sec. III.

Matched filtering. This method is known to be optimal in the case of white Gaussian noise and a known signal. If the sounds of a given species are stereotypic or nearly so, as in the case of blue whale calls (Stafford *et al.*, 1998), and if the noise environment approximates white Gaussian noise, matched filtering can be a quite effective method. In the presence of even a small amount of variation from sound to sound, or of harmonic interference such as from ship noise, matched filtering performs less well than the other methods.

Neural network. The advantage of this method is that it worked the best of the methods tried, with a best combined error rate of 1.6%. Its drawbacks are that (1) it requires a large training data set and associated operator time for data preparation, and (2) the functioning of a trained network is difficult to understand, making it uncertain under what conditions—e.g., what changes in the sound structures or noise—the network might fail. The network handled time variation in bowhead vocalizations well, probably due to its large training set.

Spectrogram correlation. This worked fairly well, with a best combined error rate of 2.5%, and should be suitable for many applications. It requires a relatively small set of training sounds—48 samples were used here, and as few as 5 samples have been used successfully in unpublished work. The shape of the kernel (center of Fig. 2) makes it easy to understand which characteristics of a sound the recognizer will and will not respond to.

ACKNOWLEDGMENTS

Thanks to the field crew in Barrow for collection of sounds, and to Steve Mitchell and Beth Weisburn for insights into the analysis. This work was supported by Grant Number N00014-93-1-0431 from the U.S. Office of Naval Research, by Contract Number N00015-94-C-6016 from the Naval Research Laboratory, by the North Slope Borough (Box 69, Barrow, Alaska), and by the National Marine Fisheries Service (NMFS) through the Alaska Eskimo Whaling Commission (AEWC) to the North Slope Borough.

Altes, R. A. (**1980**). ''Detection, estimation, and classification with spectrograms,'' J. Acoust. Soc. Am. **67**, 1232–1246.

Bradbury, J. W., and Vehrencamp, S. L. (**1998**). *Principles of Animal Communication* (Sinauer, Sunderland, MA).

Buck, J. R., and Tyack, P. L. (**1993**). ''A quantitative measure of similarity for *Tursiops truncatus* signature whistles,'' J. Acoust. Soc. Am. **94**, 2497–2506.

Chabot, D. (**1984**). ''Sound production of the humpback whale (*Megaptera novaeangliae*, Borowski) in Newfoundland waters,'' Masters thesis, Memorial University of Newfoundland, St. John's.

Chabot, D. (**1988**). ''A quantitative technique to compare and classify humpback whale (*Megaptera novaeangliae*) sounds,'' Ethology **77**, 89–102.

Charif, R. A., Mitchell, S. G., and Clark, C. W. (**1995**). *CANARY 1.2 User's Manual* (Cornell Laboratory of Ornithology, Ithaca, NY).

Clark, C. W. (**1981**). ''Acoustic Communication and Behavior of the Southern Right Whale,'' Ph.D. thesis, State University of New York at Stony Brook.

Clark, C. W. (**1982**). ''The acoustic repertoire of the southern right whale, a quantitative analysis,'' Anim. Behav. **30**, 1060–1071.
Clark, C. W. (**1983**). ''Acoustic communication and behavior of the Southern Right Whale (*Eubalaena australis*),'' in *Communication and Behavior of Whales*, edited by R. Payne (Westview, Boulder), pp. 163–198.
Clark, C. W. (**1989**). ''Call tracks of bowhead whales based on call characteristics as an independent means of determining tracking parameters,'' Rep. Intl. Whal. Commn. **39**, 111–112.
Clark, C. W. (**1990**). ''Acoustic behavior of mysticete whales,'' in *Sensory Abilities of Cetaceans*, edited by J. A. Thomas and R. A. Kastelein (Plenum, New York), pp. 571–583.
Clark, C. W. (**1991**). *Ocean Voices of the Alaskan Arctic* [audio cassette] (Cornell Laboratory of Ornithology, Ithaca, NY).
Clark, C. W., and Ellison, W. T. (**1989**). ''Numbers and distributions of bowhead whales, *Balaena mysticetus*, based on the 1986 acoustic study off Pt. Barrow, Alaska,'' Rep. Intl. Whal. Commn. **39**, 297–303.
Clark, C. W., and Ellison, W. T. (**2000**). ''Calibration and comparison of the 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.
Clark, C. W., and Fristrup, K. M. (**1996**). ''Whales '95: A combined visual and acoustic survey of blue and fin whales off Southern California,'' Rep. Intl. Whal. Commn. **47**, 583–600.
Clark, C. W., and Johnson, J. H. (**1984**). ''The sounds of the bowhead whale, *Balaena mysticetus*, during the spring migrations of 1979 and 1980,'' Can. J. Zool. **62**, 1436–1441.
Clark, C. W., and Mellinger, D. K. (**1994**). ''Application of Navy IUSS for whale research,'' J. Acoust. Soc. Am. **96**, 3315(A).
Clark, C. W., Ellison, W. T., and Beeman, K. (**1986**). ''Acoustic tracking of migrating bowhead whales,'' Proc. IEEE Oceans '86, pp. 341–346 (IEEE, New York).
Clark, C. W., Marler, P., and Beeman, K. (**1987**). ''Quantitative analysis of animal vocal phonology: An application to swamp sparrow song,'' Ethology **76**, 101–115.
Clark, C. W., Bower, J. B., and Ellison, W. T. (**1991**). ''Acoustic tracks of migrating bowhead whales, *Balaena mysticetus*, off Point Barrow, Alaska based on vocal characteristics,'' Rep. Intl. Whal. Commn. **40**, 596–597.
Clark, C. W., Mitchell, S. G., and Charif, R. A. (**1996**). ''Distribution and behavior of the bowhead whale, *Balaena mysticetus*, based on preliminary analysis of acoustic data collected during the 1993 spring migration off Point Barrow, Alaska,'' Rep. Intl. Whal. Commn. **46**, 541–554.
Clark, C. W., Tyack, P. L., and Ellison, W. T. (**1998**). Quicklook, Low-Frequency Sound Scientific Research Program. Phase I: Responses of blue and fin whales to SURTASS LFA, Southern California Bight, 5 September–21 October 1997.
Cummings, W. C., and Thompson, P. O. (**1971**). ''Underwater sounds from the blue whale, *Balaenoptera musculus*,'' J. Acoust. Soc. Am. **50**, 1193–1198.
deCharms, R. C., Blake, D. T., and Merzenich, M. M. (**1998**). ''Optimizing sound features for cortical neurons,'' Science **280**, 1439–1443.
Edds, P. L. (**1982**). ''Vocalizations of the blue whale, *Balaenoptera physalus*, in the St. Lawrence River,'' J. Mammal. **63**, 345–347.
Engineering Design. (**1997**). SIGNAL/RTS (Engineering Design, Belmont, MA).
Frankel, A. S., Clark, C. W., Herman, L. M., and Gabriele, C. M. (**1995**). ''Spatial distribution, habitat utilization, and social interactions of humpback whales, *Magaptera novaeangliae*, off Hawai'i, determined using acoustic and visual techniques,'' Can. J. Zool. **73**, 1134–1146.
Fristrup, K. M. (**1992**). ''Characterizing acoustic features of marine animal sounds,'' Woods Hole Oceanographic Institution, WHOI-92-04, Woods Hole, MA.
Fristrup, K. M., and Watkins, W. A. (**1994**). ''Marine animal sound classification,'' Woods Hole Oceanographic Institution, WHOI-94-13, Woods Hole, MA.
Gaetz, W., Jantzen, K., Weinberg, H., Spong, P., and Symonds, H. (**1993**). ''A neural network mechanism for recognition of individual Orcinus orca based on their acoustic bahavior: Phase 1,'' Proc. IEEE Oceans '93, Vol. I, pp. 455–457 (IEEE, New York).
Ghosh, J., Deuser, L. M., and Beck, S. D. (**1992**). ''A neural network-based hybrid system for detection, characterization, and classification of short-duration oceanic signals,'' IEEE J. Ocean Eng. **17**, 351–363.
Goedeking, P. (**1983**). ''A minicomputer-aided method for the detection of features from vocalizations of the cotton-top tamarin (*Saguinus oedipus oedipus*),'' Z. Tierpsychol. **62**, 321–328.
Graber, R. R. (**1968**). ''Nocturnal migration in Illinois: Different points of view,'' Wilson Bull. **80**, 36–71.
Griffin, D. R. (**1964**). *Bird Migration* (Natural History, Garden City, NY), pp. 11–13.
Hebb, D. O. (**1949**). *The Organization of Behavior* (Wiley, New York).
Herman, L. M., and Tavolga, W. N. (**1980**). ''The communication systems of cetaceans,'' in *Cetacean Behavior: Mechanism and Function*, edited by L. M. Herman (Wiley, New York), pp. 149–209.
Hubel, D. H., and Wiesel, T. N. (**1962**). ''Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex,'' J. Physiol. (London) **160**, 106–154.
Hubel, D. H., and Wiesel, T. N. (**1977**). ''Functional architecture of macaque monkey visual cortex,'' Proc. R. Soc. London, Ser. B **198**, 1–59.
Jain, A. K., and Dubes, R. C. (**1988**). *Algorithms for Clustering Data* (Prentice-Hall, Englewood Cliffs, NJ).
Kay, R. H., and Matthews, D. R. (**1972**). ''On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones,'' J. Physiol. (London) **225**, 657–677.
Larkin, R. P. (**1978**). ''Radar observations of behavior of migrating birds in response to sounds broadcast from the ground,'' in *Animal Migration, Navigation, and Homing*, edited by K. Schmidt-Koenig and W. T. Keaton (Springer, New York).
Leaper, R., Chappell, O., and Gordon, J. (**1992**). ''The development of practical techniques for surveying sperm whale populations acoustically,'' Rep. Intl. Whal. Commn. **42**, 549–560.
Lee, K.-F., Hon, H.-W., and Reddy, R. (**1990**). ''An overview of the Sphinx speech-recognition system,'' IEEE Trans. Acoust., Speech, Signal Process. **38**, 35–45.
Lippman, R. C. (**1989**). ''Review of neural networks for speech recognition,'' Neural Comput. **1**, 1–38.
McDonald, M. A., Hildebrand, J. A., and Webb, S. C. (**1995**). ''Blue and fin whales observed on a seafloor array in the Northeast Pacific,'' J. Acoust. Soc. Am. **98**, 712–721.
Mellinger, D. K. (**1993**). ''Handling time variability in bioacoustic transient detection,'' Proc. IEEE Oceans '93, pp. 116–121 (IEEE, New York).
Mellinger, D. K. (**1995**). *OSPREY 1.2 Guide*, Technical report (Cornell Laboratory of Ornithology, Ithaca, NY).
Mellinger, D. K., and Clark, C. W. (**1993**). ''A method for filtering bioacoustic transients by spectrogram image convolution,'' Proc. IEEE Oceans '93, pp. 122–127 (IEEE, New York).
Mendelson, J. R. and Cynader, M. S. (**1985**). ''Sensitivity of cat auditory primary cortex (AI) neurons to the direction and rate of frequency modulation,'' Brain Res. **327**, 331–335.
Møller, A. R. (**1977**). ''Coding of time-varying sounds in the cochlear nucleus,'' Audiology **17**, 446–468.
Moody, D. B., and Stebbins, W. C. (**1989**). ''Salience of frequency modulation in primate communication,'' in *The Comparative Psychology of Audition: Perceiving Complex Sounds*, edited by R. J. Dooling and S. H. Hulse (Erlbaum, Hillsdale, NJ).
Moore, P. W. B., Roitblat, H. L., Penner, R. H., and Nachtigall, P. E. (**1991**). ''Recognizing successive dolphin echoes with an integrator gateway network,'' Neural Networks **4**, 701–709.
Moore, S. E., Stafford, K. M., Dahlheim, M. E., Fox, C. G., Braham, H. W., and Polovina, J. J. (**1998**). ''Seasonal variation in reception of fin whale calls at five geographic areas in the North Pacific,'' Mar. Mammal Sci. **14**, 617–627.
Newman, J. D., Lieblich, A. K., Talmage-Riggs, G., and Symmes, D. (**1978**). ''Syllable classification and sequencing in twitter calls of squirrel monkeys (*Saimiri sciureus*),'' Z. Tierpsychol. **47**, 77–88.
Nyamsi, R. G. M., Aubin, T., and Bremond, J. C. (**1994**). ''On the extraction of some time-dependent parameters of an acoustic signal by means of the analytic signal concept: its application to animal sound study,'' Bioacoustics **5**, 187–203.
Oppenheim, A. V., and Schafer, R. W. (**1975**). *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).
Patrick, P. H., Ramani, N., Sheehan, R. W., and Hanson, W. (**1994**). ''Listening to and identifying wildlife using computers,'' Global Biodiversity **3**(3), 12–16.
Payne, R. S., and Payne, K. B. (**1971**). ''Underwater sounds of southern right whales,'' Zoologica **58**, 159–165.
Payne, K., Tyack, P., and Payne, R. (**1983**). ''Progressive changes in the songs of humpback whales (*Megaptera novaeangliae*): A detailed analysis of two seasons in Hawaii,'' in *Communication and Behavior of Whales*, edited by R. S. Payne (Westview, Boulder), pp. 9–57.

Payne, R. S., and McVay, S. (**1971**). ''Songs of humpback whales,'' Science **173**, 587–597.
Pinkowski, B. (**1994**). ''Robust Fourier descriptors for charactrerizing amplitude-modulated waveform shapes,'' J. Acoust. Soc. Am. **95**, 3419–3423.
Potter, J. R., Mellinger, D. K., and Clark, C. W. (**1994**). ''Marine mammal call discrimination using artificial neural networks,'' J. Acoust. Soc. Am. **96**, 1255–1262.
Rabiner, L. R., and Juang, B.-H. (**1986**). ''An introduction to hidden Markov models,'' *IEEE ASSP Magazine*, January 1986, Vol. 3, pp. 4–15.
Rabiner, L. R., and Juang, B.-H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).
Raftery, A. E., and Zeh, J. E. (**1998**). ''Estimating bowhead whale population size and rate of increase from the 1993 census,'' J. Am. Stat. Assoc. **93**, 451–463.
Ramani, N., and Patrick, P. H. (**1992**). ''Fish detection and identification using neural networks,'' IEEE J. Ocean Eng. **17**, 364–368.
Rivers, J. A. (**1997**). ''Blue whale, *Balaenoptera musculus*, vocalizations from the waters off central California,'' Mar. Mammal Sci. **13**, 186–195.
Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (**1987**). *Parallel Distributed Processing* (MIT, Cambridge, MA).
Schevill, W. E. (**1964**). ''Underwater sounds of cetaceans,'' in *Marine Bio-acoustics*, edited by W. N. Tavolga (Pergamon, New York), pp. 307–316.
Schevill, W. E., and Watkins, W. A. (**1962**). ''Whale and Porpoise Voices: A Phonograph Record,'' Contribution #1320 from Woods Hole Oceanogr. Inst., Woods Hole.
Schevill, W. E., Watkins, W. A., and Backus, R. H. (**1964**). ''The 20-cycle signals and *Balaenoptera* (fin whales),'' in *Marine Bio-acoustics*, edited by W. N. Tavolga (Pergamon, New York), pp. 147–152.
Silber, G. K. (**1986**). ''The relationship of social vocalizations to surface behavior and aggression in the Hawaiian humpback whale *Megaptera novaeangliae*,'' Can. J. Zool. **64**, 2075–2080.
Stafford, K. M., Fox, C. G., and Clark, D. S. (**1998**). ''Long-range acoustic detection and localization of blue whale calls in the northeast Pacific Ocean,'' J. Acoust. Soc. Am. **104**, 3616–3625.
Stafford, K. M., Nieukirk, S. L., and Fox, C. G. (**1999**). ''An acoustic link between blue whales in the eastern tropical Pacific and the northeast Pacific,'' J. Acoust. Soc. Am. **15**, 1258–1268.
Sturtivant, C., and Datta, S. (**1997**). ''Automatic dolphin whistle detection, extraction, encoding, and classification,'' Proc. Inst. Acoust. **19(9)**, 259–266.
Taylor, A. (**1995**). ''Bird flight call discrimination using machine learning,'' J. Acoust. Soc. Am. **97**, 3370(A).
Taylor, A., Watson, G., Grigg, G., and McCallum, H. (**1996**). ''Monitoring frog communities: an application of machine learning,'' in *Innovative Applications Artificial Intelligence Conference* (AAAI Press, Menlo Park, CA), pp. 1564–1569.
Terres, J. K. (**1980**). *The Audubon Society Encyclopedia of North American Birds* (Knopf, New York), pp. 604–605.
Thompson, P. O., Findley, L. T., Vidal, O., and Cummings, W. C. (**1996**). ''Underwater sounds of blue whales, *Balaenoptera musculus*, in the Gulf of California, Mexico,'' Mar. Mammal Sci. **12**, 288–293.
Thompson, T. J., Winn, H. E., and Perkins, P. J. (**1979**). ''Mysticete sounds,'' in *Behavior of Marine Mammals: Current Perspectives in Research*, edited by H. E. Winn and B. L. Olla (Plenum, New York), pp. 403–431.
van Trees, H. L. (**1968**). *Detection, Estimation, and Modulation Theory* (Wiley, New York), Vol. I.
Watkins, W. A., and Schevill, W. E. (**1972**). ''Sound source location by arrival-times on a non-rigid three-dimensional hydrophone array,'' Deep-Sea Res. **19**, 691–706.
Watkins, W. A., Tyack, P., Moore, K. E., and Bird, J. E. (**1987**). ''The 20-Hz signals of finback whales (*Balaenoptera physalus*),'' J. Acoust. Soc. Am. **82**, 1901–1912.
Weisburn, B. A., Mitchell, S. G., Clark, C. W., and Parks, T. W. (**1993**). ''Isolating biological acoustic transient signals,'' Proc. IEEE Intl. Conf. Acoust., Speech, Sig. Process., Vol. 1, pp. 269–272 (IEEE, New York).
Whitehead, H., and Weilgart, L. (**1990**). ''Click rates from sperm whales,'' J. Acoust. Soc. Am. **87**, 1798–1806.
Whitfield, I. C., and Evans, E. F. (**1965**). ''Responses of auditory cortical neurons to stimuli of changing frequency,'' J. Neurophysiol. **28**, 655–672.
Wiley, R. H., and Richards, D. C. (**1982**). ''Adaptations for acoustic communication in birds: Sound transmission and signal detection,'' in *Acoustic Communication in Birds*, edited by D. E. Kroodsma and E. H. Miller (Academic, London), Vol. I, pp. 131–181.
Winn, H. E., and Perkins, P. J. (**1976**). ''Distribution and sounds of the minke whale, with a review of mysticete sounds,'' Cetology **19**, 1–12.
Würsig, B., and Clark, C. W. (**1993**). ''Behavior,'' in *The Bowhead Whale*, edited by J. J. Burns, J. J. Montague, and C. J. Cowles (Allen, Lawrence, KS), pp. 157–199.
Zeh, J. E., Clark, C. W., George, J. C., Withrow, D., Carroll, G. M., and Koski, W. R. (**1993**). ''Current population size and dynamics,'' in *The Bowhead Whale*, edited by J. J. Burns, J. J. Montague, and C. J. Cowles (Allen, Lawrence, KS), pp. 409–489.