Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls

Volker B. Deecke, and Vincent M. Janik

Citation: The Journal of the Acoustical Society of America **119**, 645 (2006); doi: 10.1121/1.2139067
View online: https://doi.org/10.1121/1.2139067
View Table of Contents: https://asa.scitation.org/toc/jas/119/1
Published by the Acoustical Society of America

Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls

Volker B. Deecke
Marine Mammal Research Unit, University of British Columbia, 2202 Main Mall, Vancouver, BC V6T 1Z4 Canada and Cetacean Research Lab, Vancouver Aquarium Marine Science Centre, P.O. Box 3232, Vancouver, BC V6B 3X8, Canada

Vincent M. Janik
Sea Mammal Research Unit, Gatty Marine Laboratory, University of St. Andrews, Fife KY16 8LB, United Kingdom and Centre for Social Learning and Cognitive Evolution, School of Biology, University of St. Andrews, Fife KY16 9TS, United Kingdom

(Received 6 June 2005; revised 21 September 2005; accepted 13 October 2005)

Dividing the acoustic repertoires of animals into biologically relevant categories presents a widespread problem in the study of animal sound communication, essential to any comparison of repertoires between contexts, individuals, populations, or species. Automated procedures allow rapid, repeatable, and objective categorization, but often perform poorly at detecting biologically meaningful sound classes. Arguably this is because many automated methods fail to address the nonlinearities of animal sound perception. We present a new method of categorization that incorporates dynamic time-warping and an adaptive resonance theory (ART) neural network. This method was tested on 104 randomly chosen whistle contours from four captive bottlenose dolphins (*Tursiops truncatus*), as well as 50 frequency contours extracted from calls of transient killer whales (*Orcinus orca*). The dolphin data included known biologically meaningful categories in the form of 42 stereotyped whistles produced when each individual was isolated from its group. The automated procedure correctly grouped all but two stereotyped whistles into separate categories, thus performing as well as human observers. The categorization of killer whale calls largely corresponded to visual and aural categorizations by other researchers. These results suggest that this methodology provides a repeatable and objective means of dividing bioacoustic signals into biologically meaningful categories. [DOI: 10.1121/1.2139067]

PACS number(s): 43.80.−n, 43.60.Np, 43.80.Lb, 43.80.Ka [WWA] Pages: 645–653

I. INTRODUCTION

A. Categorization of sound patterns by humans and computers

A widespread problem in the study of animal sound communication lies in dividing the patterns that make up the acoustic repertoire of an individual or group into biologically relevant categories. We refer to this process as categorization to distinguish it from classification, the process of assigning sound patterns to predefined categories. Biologically meaningful categorization is fundamental to any study attempting to compare repertoires between contexts, individuals, populations, or species. Historically, such categorization was usually carried out by human observers who sorted the sound patterns into categories according to their perceived similarity. Categorization by human observers requires the subjects to decide which features are important in defining categories and how these features should be weighted. Humans use their natural pattern recognition skills to solve such tasks. However, the judgments and decisions made on weighting different features in a pattern can differ between individuals (Jones *et al.*, 2001) and can be difficult to quantify since humans are not usually aware of the threshold values they use (e.g., Rendell and Whitehead, 2003). This makes it difficult to compare acoustic repertoires between studies conducted by different people. One way of solving this problem is to use several observers. One can then use categories that observers agreed on and measure threshold values that distinguish these categories. However, this is a time-consuming process that limits the amount of data included in any comparison. Thus, an automated method that categorizes sound patterns in a biologically meaningful way would be an extremely valuable analytical tool.

Categorization of animal sounds has usually been based on the patterns of frequency modulation over time. Approaches to achieve automated classification have included clustering schemes based on various measures of similarity (e.g., Symmes *et al.*, 1979; Chabot, 1988), principal components analyses (e.g., Clark, 1982; Cerchio and Dahlheim, 2001), or combinations of these procedures (e.g., Nowicki and Nelson, 1990; Elowson and Hailman, 1991). However, such standard methods often fall far short of observer ratings in accuracy and frequently fail to detect biologically meaningful categories (see Janik, 1999). We argue here that this poor performance of current methods is largely due to failure to consider two fundamentals of acoustic perception when measuring the similarity of sound patterns: flexibility in the time domain and the exponential perception of sound frequency.

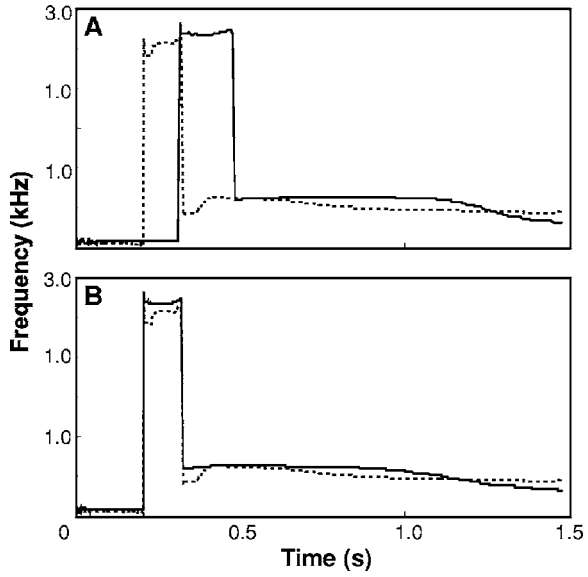

FIG. 1. Matching a frequency contour (pulse-repetition rate as a function of time) of a pulsed call of transient killer whales (solid line) to a reference contour (dotted line) using standardization of call length [panel (a)] and dynamic time warping [local extension and compression of the time axis of the frequency contour to maximize frequency overlap—panel (b)]. The match (given as the average similarity in frequency in percent for all points of the two contours) is 69.9% using standardization, but 86.9% using dynamic time warping.

B. Time and frequency resolution in the auditory perception of birds and mammals

Any categorization scheme of sound patterns requires a measure of the similarity of sounds. Traditional similarity measures have included spectrogram cross correlation (e.g., Clark *et al.*, 1987), hidden Markov models (e.g., Clemins *et al.*, 2005), or measures of the distance between frequency contours (e.g., Buck and Tyack, 1993). The first shortcoming of any study using spectrograms or frequency contours (plots of the fundamental frequency of a vocalization over time) is that, in order to compare two entire sound patterns using most traditional distance measures, they need to be standardized for time. This can have the effect of rating two sound patterns as very similar even though their lengths might differ by an order of magnitude. In addition, for signals with strong frequency modulation, temporal standardization can have the effect of generating artificially low similarity values for signals that are in fact very similar in shape, but instead differ only slightly in the length of different components so that equivalent sections of the signals do not overlap (see Fig. 1). Animals are relatively insensitive to such slight differences in the duration of sound patterns. Dooling (1982) suggests that birds are ten times more sensitive to changes in the frequency of sounds than they are to changes in their duration. Small differences in the duration of certain acoustic features are therefore often insignificant to the animal and any analysis of sound patterns must allow for this.

The other main feature of vertebrate auditory perception that needs to be considered when developing automated methods of acoustic categorization is that tonal frequency is not perceived on a linear but on an logarithmic scale. Humans perceive the difference between two tones with frequencies that differ by a factor of 2 (an octave) as being the same regardless of whether the two tones have frequencies of 110 and 220 Hz or 880 and 1760 Hz. This logarithmic perception of frequency is reflected by the distribution of hair cells sensitive to different frequencies in the inner ear and appears to be common to all terrestrial vertebrates (e.g., Müller, 1991; Smolders *et al.*, 1995; Vater and Siefer, 1995; Manley *et al.*, 1999). This means that acoustic features with higher fundamental frequencies can exhibit greater absolute frequency variation before they are perceived as different compared to features with low fundamental frequencies. Frequency measurements should therefore be log-transformed before comparison, or differences in frequencies should be expressed as relative rather than absolute values. Any scheme that fails to account for the logarithmic perception of frequency runs the risk of biasing categorization towards an inflated number of categories of high-frequency sound patterns.

C. Unsupervised learning in artificial neural networks

In this paper we introduce and test a novel method of call categorization that allows for flexibility in the time domain and accounts for the logarithmic perception of sound frequency. It uses an adaptive resonance theory (ART) neural network that is unsupervised in its learning phase. Supervised and unsupervised learning describe two different applications of self-organizing artificial neural networks. Supervised learning is a method of automated classification, where an artificial neural network learns to classify unknown patterns using information extracted from a training set of identified patterns. For example, artificial neural networks can be trained in this way to distinguish between the vocal patterns of different identified individuals (e.g., Reby *et al.*, 1997; Campbell *et al.*, 2002; Terry and McGregor, 2002), social groups (e.g., Deecke *et al.*, 1999), or species (e.g., Phelps and Ryan, 1998; Parsons and Jones, 2000), or between vocal patterns given in response to clearly identifiable stimuli [e.g., predator-specific calls (Placer and Slobodchikoff, 2000)]. In contrast, unsupervised learning describes a series of artificial neural network algorithms that can be used to categorize patterns without prior training—they are the self-organizing analogs of traditional clustering schemes. The main advantage of unsupervised learning is that, for a new pattern to be assigned to a category, it must only be compared to a small subset of reference patterns (or neighboring patterns in the case of self-organizing maps) rather than all other patterns in the data set. Unsupervised learning algorithms therefore lend themselves to the analysis of large data sets where computing time is limiting, or to situations where categorization must happen in real time.

The most common algorithms for unsupervised learning are self-organizing maps [SOM, e.g., Kohonen (1988)], competitive learning (e.g., Grossberg, 1987), and adaptive resonance theory (ART) neural networks (e.g., Carpenter and Grossberg, 1987). The categorization algorithm used in this study is based on an ART2 neural network (Carpenter and Grossberg, 1987). ART2 is an unsupervised learning algorithm in which a given pattern is compared to a set of refer-

ence patterns. If the input pattern resembles one of the reference patterns with a certain degree of similarity (called the vigilance), the input is assigned to the category represented by this reference pattern and the reference pattern itself is updated and made even more similar to the current input pattern. If the input pattern does not resemble any reference pattern sufficiently, it becomes the reference pattern for a new category. ART2 neural networks have the advantage that they do not require assumptions about the frequencies of patterns in different categories. In contrast, competitive learning algorithms and self-organizing maps assume that input patterns are evenly distributed between categories and therefore tend to split frequent input patterns into finer categories. For this reason, ART neural networks lend themselves to the categorization of behavior patterns where equal distribution can rarely be assumed.

D. Objectives

Our objective for this study is to develop and test an automated method for categorizing stereotyped vocal patterns of animals using test data sets of vocalizations of bottlenose dolphins and killer whales. Both of these species produce a variety of structurally distinct stereotyped sound patterns and dividing these into meaningful sound categories is an important first step before vocal repertoires, or the structure of given sound types, can be compared between individuals and populations, behavior contexts, or over time. While the methodology is developed using data sets from two species of toothed whales, our hope is that it can be applied to the vocalizations of a wide variety of species.

In order to allow for variation in the lengths of different components of the sounds, similarities between input and reference patterns were calculated using dynamic time-warping (e.g., Sakoe and Chiba, 1978; Buck and Tyack, 1993). Dynamic time-warping is an algorithm developed for the automated recognition of human speech that allows limited compression and expansion of the time axis of a signal to maximize frequency overlap with a reference signal (see Fig. 1 for an illustration of dynamic time-warping). To account for exponential perception of frequency in this analysis, we expressed similarity of contours as their relative similarity in frequency.

We test the performance of this method on two categorization problems. The first is a set of frequency contours of bottlenose dolphin whistles described in detail by Janik (1999). Bottlenose dolphins produce a variety of whistles, including stereotyped signature whistles which are individually distinctive. Since signature whistles represent biologically defined categories and their structure has been shown to convey important information (i.e., identity of the caller) to the animals (Janik and Slater, 1998; Sayigh *et al.*, 1999), we aim to use this data set to test whether the categories determined by our analysis are congruent with categories known to be perceived as meaningful by bottlenose dolphins.

The second data set consists of frequency contours of killer whale calls. The pulsed calls of killer whales are highly stereotyped and can be divided easily and consistently into categories by human observers (e.g., Ford, 1989, 1991). We present the categorization performance and investigate the vigilance parameter that controls the fineness of categorization and therefore the number of categories established. We also show how optimality of categorization can be achieved without prior knowledge of the underlying categories by selecting a vigilance parameter for the neural network that minimizes variation within categories while maximizing differences between categories.

II. METHODS

A. Data sets, acoustic analysis, and contour extraction

Both the dolphin whistle and killer whale call data sets consist of frequency contours extracted from spectrograms of calls or whistles. Dolphin whistles are tonal signals and frequency contours therefore give the fundamental frequency of a whistle as a function of time. The recordings for our study were collected in 1996 from a social group consisting of two female and two male bottlenose dolphins held at Zoo Duisburg in Germany. We recorded whistles with two Dowty SSQ 904 hydrophones connected to a Marantz CP430 tape recorder (system frequency response: 1–20 kHz±3 dB). Time resolution of the extracted frequency contours was 10 ms. For details on the recording and selection of bottlenose dolphin whistles and on the extraction of frequency contours see Janik *et al.* (1994); Janik and Slater (1998), and Janik (1999).

The frequency contours of killer whale calls were generated from a sample of calls digitized from field recordings of transient killer whales. Recordings were made with a variety of hydrophones on Type II audio cassette tapes, digital audio tape, or reel-to-reel tape. Frequency responses of the recording systems were 1–16 kHz±3 dB or better. We rated the quality of each call from the spectrogram on a scale from one to five, taking into account signal-to-noise ratio, echoes, and reverberation, as well as background noise. In order to avoid categorization due to noise artifacts (e.g., faint call elements that were missed), calls of the three lowest quality categories were excluded from this analysis. Since killer whale calls are pulsed signals (Schevill and Watkins, 1966), frequency contours give the pulse-repetition rate rather than fundamental frequency. We used the sidewinder algorithm (Deecke *et al.*, 1999) to extract frequency contours from spectrograms of killer whale calls, with the difference that for the current analysis the contours were not standardized for time. Time resolution for these frequency contours was also 10 ms.

B. ARTwarp—Combining dynamic time-warping and adaptive resonance theory

The neural network used in this analysis was an ART2 neural network for the categorization of analog input patterns. The computer script was a simulation of the ART2 algorithm of Carpenter and Grossberg (1987). However, this algorithm was modified in two ways. First, the similarity between frequency contours and the set of reference contours was calculated using dynamic time-warping to ensure maximum overlap in the frequency domain. Second, if a fre-

quency contour matched a reference contour better than the critical similarity (vigilance), this reference contour was then modified in three ways to be more similar to the input pattern. (1) The frequency content of the reference contour was made more similar to the time-warped frequency contour by adding a proportion (10%) of the difference between reference contour and time-warped input contour. (2) The relative lengths of different components of the reference contour were modified to be more similar to the current frequency contour by applying a warping function that stretched or compressed the time axis by a proportion (10%) of the inverse of the original warping function generated when input and reference contour were compared. (3) The length of the reference contour was made more similar to the current input contour. The extent of the change in length (number of points) was given by the learning rate (10% of the difference in our case). To increase or decrease the number of frequency points, the frequency measurements in the contour were interpolated at a number of equally spaced points corresponding to the new length of the contour. If the current input pattern did not match any of the reference patterns better than the critical similarity, it became the reference contour for a new category. All frequency contours were repeatedly presented to the neural network until they consistently matched the same reference contour (i.e., no reclassifications occurred between iterations).

The dynamic time-warping algorithm used in this study was that applied by Sakoe and Chiba (1978) and Buck and Tyack (1993) with the difference that the algorithm allowed horizontal and vertical jumps of three frequency points in the contour [rather than two points as in Sakoe and Chiba (1978) and Buck and Tyack (1993)]. A frequency contour can therefore be "sped up" or "slowed down" in parts by a factor of 3 to fit the reference contour. In addition, the algorithm calculated the relative frequency similarity (S) in percent between both frequency contours rather than the total square difference of Sakoe and Chiba (1978) or the average square difference of Buck and Tyack (1993). This was done by dividing the smaller frequency value by the larger value at each point and multiplying by 100:

$$S(i) = \frac{\min[M(i),N(i)]}{\max[M(i),N(i)]} \times 100,$$

where M is the reference pattern and N the input pattern. Like Buck and Tyack (1993), we also divided the total difference by the length of the reference contour. The measure of similarity therefore gives the average relative similarity in frequency for the reference and input contour after time warping.

C. Experiment I: Categorization of bottlenose dolphin whistles

The level of critical similarity (vigilance) for the analysis of dolphin whistles was obtained by categorizing only the signature whistles of one individual [individual A of Janik (1999)] and increasing the vigilance in steps of 1% until the analysis split these signature whistles into two categories. The critical vigilance (96%) is the highest value that still recognizes the whistles as a single category. The entire data set was then categorized using this vigilance parameter and the resulting categories were analyzed to test whether the signature whistle categories were recognized.

D. Experiment II: The appropriate fineness of categorization

In this experiment, we categorized a sample of 50 frequency contours randomly chosen from all calls with the two highest quality scores in the transient killer whale data set. These calls came from 25 recordings of different groups. Unlike the bottlenose dolphin whistles, this data set does not contain any sound categories known *a priori* to be biologically meaningful. Therefore the method to determine the appropriate fineness of categorization used for the dolphin whistles could not be applied. In the absence of identifiable categories, we wanted to find the categorization that would explain a maximum of the variation in call structure with a minimum number of call categories. To do this, we initially set the vigilance to zero. At this level, call categories are assigned only by call length [since any two contours whose length differs by more than a factor of 3 are automatically assigned a similarity of zero; see Buck and Tyack (1993)]. The vigilance was then increased to 100% in 50 logarithmic steps and the sample of contours was categorized. At a vigilance of 100%, each frequency contour is assigned to its own category. For each categorization, we determined the number of categories generated. In order to investigate patterns of within-category and between-category variation, we calculated the similarity matrix for all frequency contours in the data. This matrix contained similarity values of all possible pairwise comparisons of contours using dynamic time-warping. Using this matrix, we could determine the average similarity of contours in the same category (within-category variation), as well as the average similarity of contours in different categories (between-category variation) for each categorization. The categorization where a minimum number of distinct categories explain a maximum amount of difference in the frequency contours can then be identified by plotting the ratio of within-category and between-category variation and determining the number of categories where this ratio levels off (i.e., adding additional categories does little to explain additional variation). This is analogous to the variance ratio criterion (e.g., Calinski and Harabasz, 1974; Everitt *et al.*, 2001; Schreer *et al.*, 1998; Rendell and Whitehead, 2003) to determine the optimal number of groups in cluster analysis.

E. Visualization of neural network performance

In order to illustrate how the ARTwarp algorithm categorizes the discrete calls of killer whales from frequency contours, we used the neural network to categorize a sample of 20 frequency contours, small enough so that it could be graphed on a single page. These were randomly chosen from the two highest quality categories in the data set of transient killer whale calls. The vigilance parameter used in this analysis was the optimum value determined in experiment II.

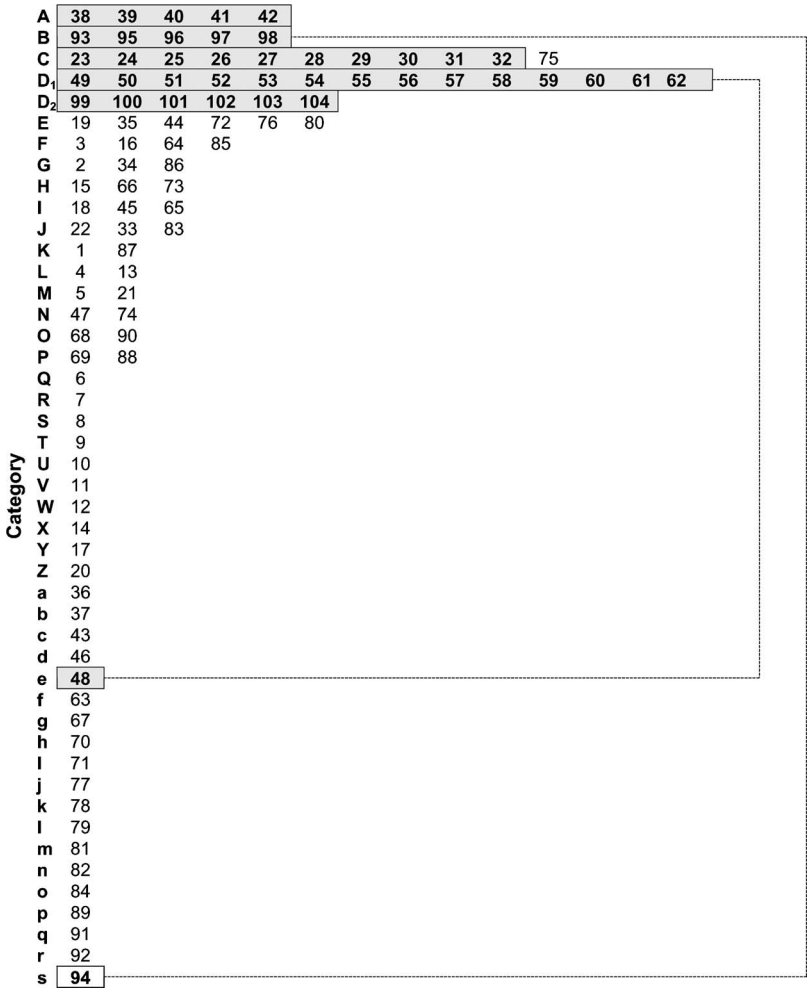

FIG. 2. Categorization of frequency contours of bottlenose dolphin whistles using an ART2 neural network and dynamic time warping to calculate similarity. Numbers represent individual whistle contours. Signature whistles are shown in bold and boxes identify signature whistles from the same individual. Signature whistle categories that were split by the analysis are linked with dotted lines. See Janik (1999) for visual representations of the whistle contours.

III. RESULTS

A. Experiment I: Categorization of bottlenose dolphin whistles

The categorization of the data set of bottlenose dolphin whistles is shown in Fig. 2. Using a vigilance of 96%, the analysis divided the 104 whistle contours into 46 categories each containing between 1 and 14 contours (mean: 2.26, standard deviation: 2.62 contours). With regard to the behaviorally defined categories of signature whistles recorded from each of the five dolphins in isolation, the analysis correctly clustered two whistle types (A and D2) into individual categories but made three errors while categorizing the other three whistle types: It added an additional whistle (no. 75) to the category containing the contours of whistle type C. In the case of whistle types B and D_1, a single contour was not assigned to the category containing the whistle types, but was put in a category of its own.

B. Experiment II: The appropriate fineness of categorization

The effects of increasing the vigilance parameter on the categorization of transient killer whale calls are illustrated in Fig. 3. With higher vigilance the analysis generated an increasing number of categories. Both the average similarity of frequency contours within a category and the average similarity of contours in different categories increased as more and more categories were added. However, they did so at different rates. Initially the rate of increase in the between-category similarity was relatively low and the rate of increase in the within category variation was high. At a critical point, however, the rate of increase in the within-category similarity slowed (since new categories explain little additional variation) and the rate of increase in the between-category similarity increased (since more and more natural clusters in the data set were divided between categories). Adding further categories after this critical point does little to improve the categorization. The plot of the ratio of within and between-category variation [Fig. 3(b)] therefore leveled off abruptly at a vigilance of 81.24%. At this point the analysis generated ten categories.

C. Visualization of neural network performance

The frequency contours used in this experiment, as well as the resulting call categories, are shown in Fig. 4. The analysis divided the 20 contours into six categories each containing between two and seven calls. The categories were largely consistent with the call types established by Ford (1984) and Ford and Morton (1991): Category 1 contained calls classified as T08i, category 2 represents the T04 call type of Ford (1984) and the T03ii call type of Ford and Morton (1991), category 3 is equivalent to the T01 call type,

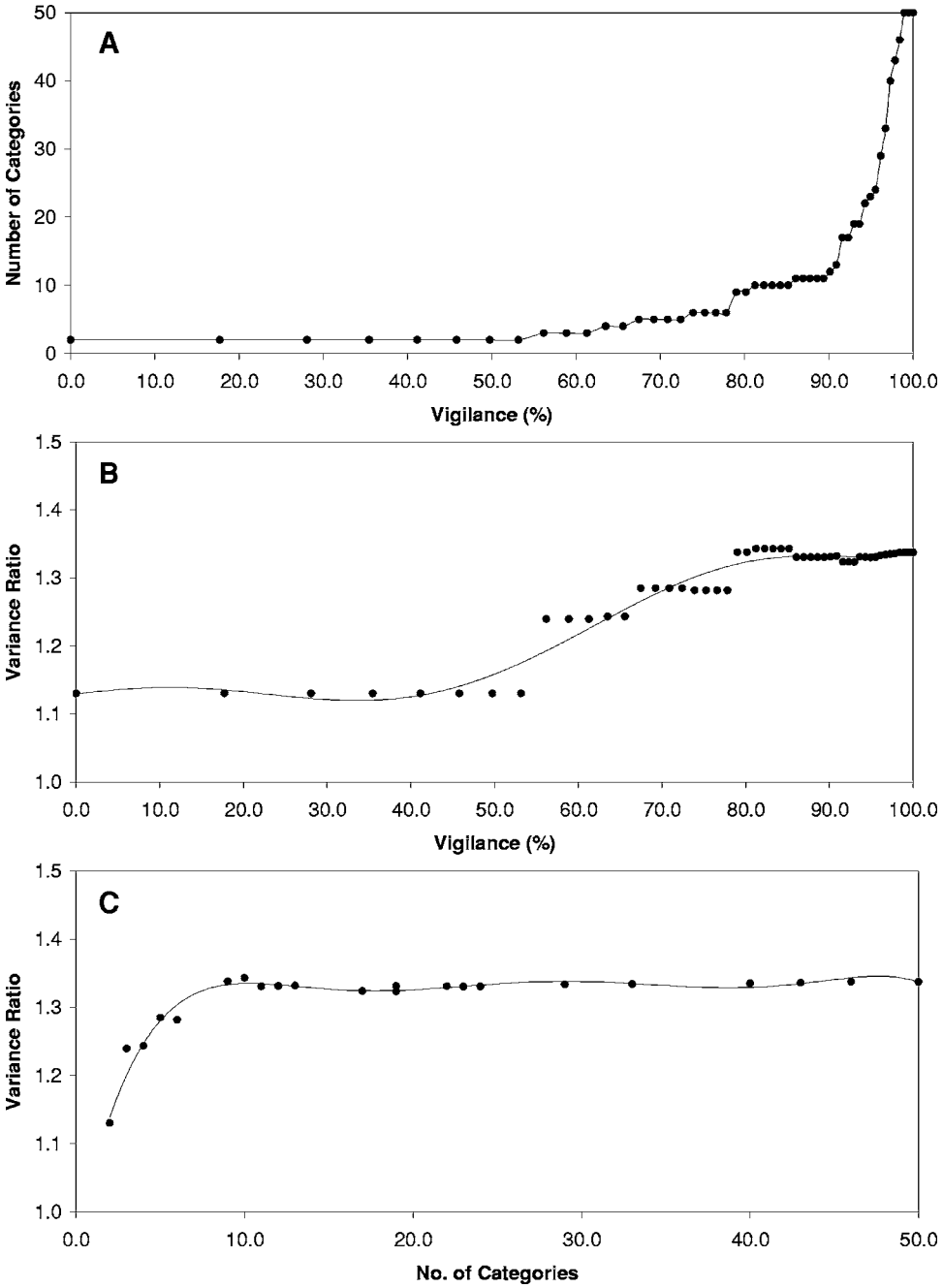

FIG. 3. Effect of the vigilance on the categorization of 50 frequency contours from calls of transient killer whales. Panel (a) shows the increase in the number of categories generated with increasing vigilance. Panel (b) shows the change in the variance ratio (ratio of within- to between-category variance) with increasing vigilance and panel (c) shows the change in the variance ratio with increasing numbers of categories. This ratio reached a maximum at a vigilance of 81.24% (10 categories). Trend lines in panels (b) and (c) are sixth-order polynomials.

and category 4 represents the T07 call type of Ford and Morton (1991). Category 5 contained calls classified as subtype T07ii by Ford and Morton (1991) and category 6 is equivalent to their T02 call type.

IV. DISCUSSION

A. Categorization of bottlenose dolphin whistles

The automated categorization combining dynamic time-warping with an ART2 neural network was able to recognize biologically meaningful categories in our data set of bottlenose dolphin whistles. While the analysis was not designed to detect individual signature whistles and identify them as such (a problem of *classification*, not categorization) it did recognize the stereotyped signature whistles as uniform signal categories to a high degree. By doing so, our method performed much better than any of the statistical procedures tested by Janik (1999), none of which proved satisfactory at detecting these biologically significant signal categories. Our automated categorization even performed marginally better at detecting the signature whistle categories in the data set of bottlenose dolphin whistles than did the human observers of Janik (1999) who made on average 3.4 mistakes. Interestingly, the neural network did not agree with the human observers in the categorization of nonsignature whistles. In general, the automated analysis created finer categories containing fewer contours for this subset. Janik (1999) identified four combinations of nonsignature whistles common to the categorization of all five observers and none of these combinations occur in the neural network categorization. Since we have no external validation for appropriate classification of nonsignature whistles, it is impossible to say which catego-

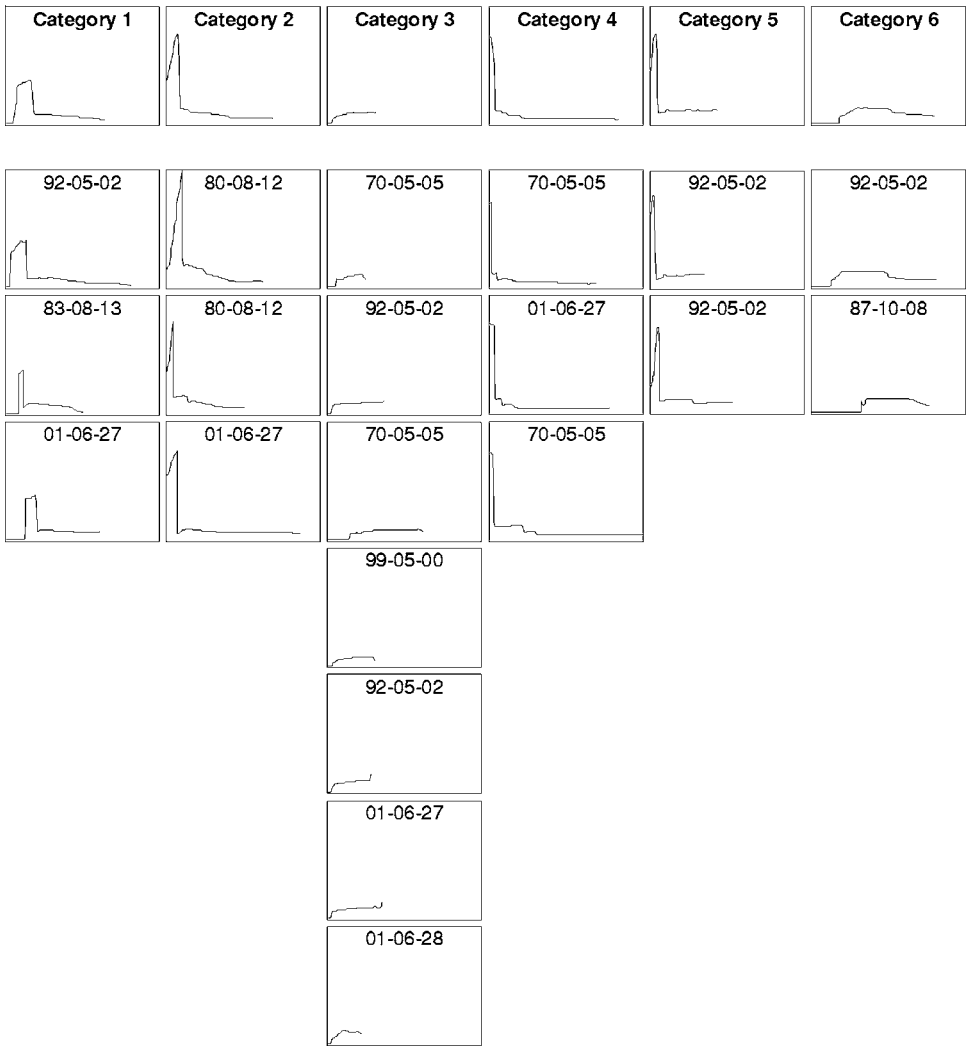

FIG. 4. Results of the categorization of frequency contours from 20 randomly chosen calls of transient killer whales. All frequency contours in the same column were assigned to the same call type by the analysis. The reference contours representing each category are shown in the first row. Labels give the recording session (in the format yy-mm-dd) for each frequency contour.

rization scheme is of greater biological relevance here.

The two signature whistles that were assigned to separate categories from the rest of their whistle types are both shorter than the other whistles of the same type and may represent truncated versions of the individuals' signature whistles. If this is the case, relaxing the endpoint constraint during dynamic time-warping [i.e., permitting the time-warped contour to be shorter in duration than the reference contour and calculating frequency similarity only for the section of overlap with the reference contour; see Parsons (1987)] would improve classification for these contours.

B. Choosing the vigilance parameter

Most automated analytic procedures require the investigator to choose some parameters that control their performance. In the automated categorization described here, the performance depends to a large degree on the vigilance of the neural network. This parameter controls the fineness of categorization, that is, the size and number of categories that are generated. It has no influence on which patterns are rated as similar in the analysis. Note that the problem of deciding on the appropriate fineness of categorization is shared by categorization of behavior using human observers: we refer to observers as "joiners" or "splitters" depending on how fine their behavior categories tend to be. As an example, Saulitis (1993) divides the surface behavior of killer whales into 14 categories, whereas Ford (1989) distinguishes between only five behavior categories. We have no information on the extent to which this difference is due to differences in the behavior of killer whale populations studied by the two researchers, or differences in the fineness of categorization considered appropriate to describe the observed behavioral variation by the authors. The advantage of the automated procedure is obviously that the fineness of categorization can be quantified for each analysis.

The categorization of bottlenose dolphin whistles demonstrates that where biologically relevant sound categories can be identified *a priori*, these can be used to determine the vigilance parameter appropriate for categorization. Such biologically defined categories may be sound patterns specific to certain individuals or populations or to clearly defined contexts, such as isolation from group members (Symmes *et al.*, 1979; Janik, 1999) or the presence of a food source (Judd and Sherman, 1996; Roush and Snowdon, 2000) or a predator (e.g., Placer and Slobodchikoff, 2000). Human observers frequently use such information from predefined categories to determine the appropriate resolution for behavioral categorization.

In many categorization problems, it is desirable to ex-

plain a maximum amount of the observed acoustic variation using a minimum number of sound categories. In situations where acoustic variation is difficult to quantify, this can be hard to achieve. However, wherever measures of acoustic similarity are readily available, simple algorithms can help to determine the appropriate number of categories for analysis. In situations where no external validation of categories is available, calculating the ratio of variation within to variation between categories for a large number of vigilance values provides a useful approach to determining the appropriate fineness of categorization. This is time consuming for large samples of sound patterns but, as demonstrated in experiment II, categorization of a randomly selected subset will generally allow identification of the appropriate vigilance parameter. Applying alternative goodness of fit measures, such as the Bayesian information criterion or Aikake's information criterion (e.g., Kuha, 2004), to categorizations with increasing vigilance setting may similarly help to identify the appropriate fineness of categorization in future studies.

C. Applicability to other categorization problems in the study of behavior

While this method has so far only been tested on the vocalizations of toothed whales, these results should also encourage its application to analyses of vocal behavior in other species. Unsupervised learning algorithms have been applied successfully to the categorization and classification of a variety of bioacoustic signals (e.g. Leinonen *et al.*, 1993; Terry and McGregor, 2002) and allowing for differences in the length of acoustic signals and their components by incorporating dynamic time-warping may prove useful in these and other situations as well. As described here, our analysis is currently limited to vocalizations that can be described adequately by frequency contours. This includes the sound signals of many species of amphibians, birds, and mammals. However, in species with vocalizations that are broadband (e.g., Campbell *et al.*, 2002), or where relevant information is encoded in the harmonic content (e.g., Weiss and Hauser, 2002), frequency contours alone are inadequate to describe vocal patterns. Fortunately, dynamic time-warping can also be used to compare spectrograms [it was in fact first developed to classify human speech patterns from spectrograms (see Sakoe and Chiba, 1978)] and the neural network component of the analysis could easily be adapted to deal with the two-dimensional format of spectrograms rather than one-dimensional frequency contours, making the analysis applicable to the categorization of vocal behavior in a wide variety of species.

Since it was developed to address peculiarities of acoustic perception, the methodology as described in this study is probably of limited value to categorize behaviors other than those that are acoustic. Nonetheless, elements applied in the current analysis may prove useful elsewhere: the ART2 neural network can be used with similarity measures other than dynamic time-warping in a wide variety of categorization problems. Conversely, dynamic time-warping and its extension of hidden Markov models will be useful in any situation where the trajectory of change in a behavioral parameter is more important than the precise timing of the change. The categorization of dive profiles from aquatic birds and mammals (e.g., Schreer *et al.*, 1998; Lesage *et al.*, 1999; Malcolm and Duffus, 2000) may prove to be a valuable example. In addition, much if not most of sensory perception is nonlinear in scale (usually exponential or logarithmic), and this is important to bear in mind when quantifying the strength and quality of behavioral stimuli for categorization. This study therefore serves to illustrate the importance of obtaining and applying relevant information about the sensory perception of study animals when designing categorization schemes for the study of their behavior.

V. CONCLUSIONS

Our results suggest that automated categorization of bioacoustic signals can present a powerful alternative to categorization by human observers, as long as the importance of the time domain and the frequency domain in the auditory perception of the study species is understood and any peculiarities in the way time and frequency parameters are perceived are considered. Automated methods such as the one used in our study are particularly useful in situations were large data sets need to be analyzed and where the size of acoustic repertoires must be compared between individuals, social groups or species, or over time.

ACKNOWLEDGMENTS

We thank the staff of Zoo Duisburg for the opportunity to work with their animals and for their support during the recording of dolphin whistles, especially Roland Edler, Reinhard Frese, Manuel García Hartmann, Friedrich Ostenrath, and Ulf Schönfeld. Recordings for the analysis of killer whale vocalizations were generously supplied by Nancy A. Black, John K. B. Ford, P. Dawn Goley, Dan McSweeney, Paul Spong, and Richard L. Ternullo. The ART2 neural network algorithm was adapted from a program originally written by Aaron Garrett, and Mary Royerr helped with statistical aspects of this paper. Earlier drafts of this manuscript benefited from comments by Karen E. McComb, John K. B. Ford, Michael J. Ritchie, and Peter J. B. Slater. V.M.J was funded by a Royal Society University Research Fellowship, and V.B.D received financial support from a DAAD Doktorandenstipendium aus Mitteln des 3. Hochschulsonderprogramms during part of this study.

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.

Calinski, T., and Harabasz, J. (**1974**). "A dendrite method for cluster analysis." Commun. Stat: Theory Meth. **3**, 1–27.

Campbell, G. S., Gisiner, R. C., Helweg, D. A., and Milette, L. L. (**2002**). "Acoustic identification of female Steller sea lions (*Eumetopias jubatus*)," J. Acoust. Soc. Am. **111**, 2920–2928.

Carpenter, G. A., and Grossberg, S. (**1987**). "ART 2: Self-organization of stable category recognition codes for analog input patterns," Appl. Opt. **26**, 4919–4930.

Cerchio, S., and Dahlheim, M. E. (**2001**). "Variation in feeding vocalizations of humpback whales (*Megaptera novaeangliae*) from Southeast Alaska," Bioacoustics 11, 277–295.

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

Clark, C. W. (**1982**). "The acoustic repertoire of the southern right whale, a quantitative analysis," Anim. Behav. **30**, 1060–1071.

Clark, C. W., Marler, P., and Beeman, B. (**1987**). "Qualitative analysis of animal vocal phonology and application to swamp sparrow song," Ethology **76**, 101–115.
Clemins, P. J., Johnson, M. T., Leong, K. M., and Savage, A. (**2005**). "Automatic classification and speaker identification of African elephant (*Loxodonta africana*) vocalizations," J. Acoust. Soc. Am. **117**, 956–963.
Deecke, V. B., Ford, J. K. B., and Spong, P. (**1999**). "Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale (*Orcinus orca*) dialects," J. Acoust. Soc. Am. **105**, 2499–2507.
Dooling, R. J. (**1982**). "Auditory perception in birds," in *Acoustic Communication in Birds*, edited by E. D. Kroodsma, E. H. Miller, and H. Ouellet (Academic, London), pp. 95–131.
Elowson, A. M., and Hailman, J. P. (**1991**). "Analysis of complex variation: Dichotomous sorting of predator-elicited calls of the Florida scrub jay," Bioacoustics 3, 295–320.
Everitt, B. S., Landau, S., and Leese, M. (**2001**). *Cluster Analysis*, 4th ed. (Arnold, London), pp. 102–105.
Ford, J. K. B. (**1984**). "Call Traditions and Vocal Dialects of Killer Whales (*Orcinus orca*) in British Columbia," (Ph.D. dissertation, University of British Columbia, Vancouver, BC).
Ford, J. K. B. (**1989**). "Acoustic behaviour of resident killer whales (*Orcinus orca*) off Vancouver Island, British Columbia," Can. J. Zool. **67**, 727–745.
Ford, J. K. B. (**1991**). "Vocal traditions among resident killer whales (*Orcinus orca*) in coastal waters of British Columbia, Canada," Can. J. Zool. **69**, 1454–1483.
Ford, J. K. B., and Morton, A. B. (**1991**). "Vocal behaviour and dialects of transient killer whales in coastal waters of British Columbia, California and southeast Alaska," in *Abstracts of the Ninth Biennial Conference on the Biology of Marine Mammals* (Society for Marine Mammalogy, Chicago, IL).
Grossberg, S. (**1987**). "Competitive learning: From interactive activation to adaptive resonance," Cogn. Sci. **11**, 23–63.
Janik, V. M. (**1999**). "Pitfalls in the categorization of behaviour: A comparison of dolphin whistle classification methods," Anim. Behav. **57**, 133–143.
Janik, V. M., and Slater, P. J. B. (**1998**). "Context-specific use suggests that bottlenose dolphin signature whistles are cohesion calls," Anim. Behav. **56**, 829–838; Printer's Erratum: Anim. Behav.57, 1173.
Janik, V. M., Dehnhardt, G., and Todt, D. (**1994**). "Signature whistle variations in a bottlenosed dolphin, *Tursiops truncatus*," Behav. Ecol. Sociobiol. **35**, 243–248.
Jones, A. E., Ten Cate, C., and Bijleveld, C. J. H. (**2001**). "The interobserver reliability of scoring sonagrams by eye: A study on methods, illustrated on zebra finch songs," Anim. Behav. **62**, 791–801.
Judd, T. M., and Sherman, P. W. (**1996**). "Naked mole-rats recruit colony mates to food sources," Anim. Behav. **52**, 957–969.
Kohonen, T. (**1988**). "The self-organizing map," Proc. IEEE **78**, 1464–1480.
Kuha, J. (**2004**). "AIC and BIC: Comparisons of assumptions and performance," Sociolog. Methods Res. **33**, 188–229.
Leinonen, L., Hiltunen, T., Torkkola, K., and Kangas, J. (**1993**). "Self-organized acoustic feature map in detection of fricative-vowel coarticulation," J. Acoust. Soc. Am. **93**, 3468–3472.
Lesage, V., Hammill, M. O., and Kovacs, K. M. (**1999**). "Functional classification of harbor seal (*Phoca vitulina*) dives using depth profiles, swimming velocity, and an index of foraging success," Can. J. Zool. **77**, 74–87.
Malcolm, C. D., and Duffus, D. A. (**2000**). "Comparison of subjective and statistical methods of dive classification using data from a time-depth recorder attached to a gray whale (*Eschrichtius robustus*)," J. Cetacean Res. Manage. **2**, 177–182.
Manley, G. A., Koppl, C., and Sneary, M. (**1999**). "Reversed tonotopic map of the basilar papilla in *Gekko gecko*," Hear. Res. **131**, 107–116.
Müller, M. (**1991**). "Frequency representation in the rat cochlea," Hear. Res. **51**, 247–254.
Nowicki, S., and Nelson, D. A. (**1990**). "Defining natural categories in acoustic signals: comparison of three methods applied to 'chick-a-dee' call notes," Ethology **86**, 89–101.
Parsons, S., and Jones, G. (**2000**). "Acoustic identification of twelve species of echolocating bat by discriminant function analysis and artificial neural networks," J. Exp. Biol. **203**, 2641–2656.
Parsons, T. W. (**1987**). *Voice and Speech Processing* (McGraw-Hill, New York).
Phelps, S. M., and Ryan, M. J. (**1998**). "Neural networks predict response biases of female túngara frogs," Proc. R. Soc. London, Ser. B **265**, 279–285.
Placer, J., and Slobodchikoff, C. N. (**2000**). "A fuzzy-neural system for identification of species-specific alarm calls of Gunnison's prairie dogs," Behav. Processes **52**, 1–9.
Reby, D., Lek, S., Dimopoulos, I., Joachim, J., Lauga, J., and Aulagnier, S. (**1997**). "Artificial neural networks as a classification method in the behavioural sciences," Behav. Processes **40**, 35–43.
Rendell, L. E., and Whitehead, H. (**2003**). "Comparing repertoires of sperm whale codas: A multiple methods approach," Bioacoustics 14, 61–81.
Roush, R. S., and Snowdon, C. T. (**2000**). "Quality, quantity, distribution and audience effects on food calling in cotton-top tamarins," Ethology **106**, 673–690.
Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-26**, 43–49.
Saulitis, E. L. (**1993**). "The Behavior and Vocalizations of the 'AT' Group of Killer Whales (*Orcinus orca*) in Prince William Sound, Alaska," (M.Sc. thesis, University of Alaska, Fairbanks, AK).
Sayigh, L. S., Tyack, P. L., Wells, R. S., Solow, A. R., Scott, M. D., and Irvine, A. B. (**1999**). "Individual recognition in wild bottlenose dolphins: A field test using playback experiments," Anim. Behav. **57**, 41–50.
Schevill, W. E., and Watkins, W. A. (**1966**). "Sound structure and directionality in *Orcinus* (killer whale)," Zoologica (N.Y.) (N.Y.) **51**, 70–76.
Schreer, J. F., Hines, R. J. O., and Kovacs, K. M. (**1998**). "Classification of dive profiles: A comparison of statistical clustering techniques and unsupervised artificial neural networks," J. Agric. Biol. Environ. Stat. **3**, 383–404.
Smolders, J. W. T., Ding-Pfennigdorff, D., and Klinke, R. (**1995**). "A functional map of the pigeon basilar papilla: Correlation of the properties of single auditory nerve fibres and their peripheral origin," Hear. Res. **92**, 151–169.
Symmes, D., Newman, J. D., Talmage-Riggs, G., and Katz Lieblich, A. (**1979**). "Individuality and stability of isolation peeps in squirrel monkeys," Anim. Behav. **27**, 1142–1152.
Terry, A. M. R., and McGregor, P. K. (**2002**). "Census and monitoring based on individually identifiable vocalizations: The role of neural networks," Animal Conservation **5**, 103–111.
Vater, M., and Siefer, W. (**1995**). "The cochlea of *Tadarida brasiliensis*: Specialized functional organization in a generalized bat," Hear. Res. **91**, 178–195.
Weiss, D. J., and Hauser, M. D. (**2002**). "Perception of harmonics in the combination long call of cottontop tamarins, *Sanguinus oedipus*," Anim. Behav. **64**, 415–426.