Taylor & Francis
Taylor & Francis Group

Bioacoustics
The International Journal of Animal Sound and its Recording

ISSN: 0952-4622 (Print) 2165-0586 (Online) Journal homepage: http://www.tandfonline.com/loi/tbio20

COMPARING REPERTOIRES OF SPERM WHALE CODAS: A MULTIPLE METHODS APPROACH

L. E. RENDELL & H. WHITEHEAD

To cite this article: L. E. RENDELL & H. WHITEHEAD (2003) COMPARING REPERTOIRES OF SPERM WHALE CODAS: A MULTIPLE METHODS APPROACH, Bioacoustics, 14:1, 61-81, DOI: 10.1080/09524622.2003.9753513

To link to this article: https://doi.org/10.1080/09524622.2003.9753513

Bioacoustics
The International Journal of Animal Sound and its Recording, 2003, Vol. 14, pp. 61–81
0952-4622/03 $10

COMPARING REPERTOIRES OF SPERM WHALE CODAS: A MULTIPLE METHODS APPROACH

L. E. RENDELL* AND H. WHITEHEAD

Department of Biology, Dalhousie University, Halifax, Nova Scotia, B3H 4J1, Canada

ABSTRACT

A common task for researchers of animal vocalisations is statistically comparing repertoires, or sets of vocalisations. We evaluated five methods of comparing repertoires of 'codas', short repeated patterns of clicks, recorded from sperm whale (*Physeter macrocephalus*) groups. Three of the methods involved classification of codas – human observer classification, k-means cluster analysis using Calinski and Harabasz's (1974) criterion to determine k, and a divisive k-means clustering procedure using Duda and Hart's (1973) criterion to determine k. Two other methods used multivariate distances to calculate similarity measures between coda repertoires. When used on a sample coda dataset, observer classification failed to produce consistent results. Calinski and Harabasz's criterion did not provide a clear signal for determining the number of coda classes (k). Divisive clustering using Duda and Hart's criterion performed satisfactorily and, encouragingly, gave similar results to the multivariate similarity measures when used on our data. However, the relative performance of the k-means techniques is likely data dependent, so one method is not likely to perform best in all circumstances. Thus results should be checked to ensure they extract logical clusters. Using these techniques concurrently with multivariate measures allows the drawing of relatively robust conclusions about repertoire similarity while minimising uncertainties due to questionable validity of classifications.

Keywords: cluster analysis, classification, vocal repertoire, sperm whale, codas

INTRODUCTION

In the study of animal vocalisations, the problem of objectively defining categories and statistically comparing repertoires between individuals or sets of animals is perennial (see for example Janik (1999); Nowicki & Nelson (1990); Terhune et al. (1993)). Here we describe and compare a number of methods that we have developed to study the repertoires of 'coda' vocalisations in sperm whale (*Physeter macrocephalus*) social groups. Codas are repeated stereotyped

*Corresponding author. Email: lrendell@dal.ca

sequences of 3-40 broadband (0-16 kHz) clicks generally heard during periods of socialising (Watkins & Schevill 1977). Sperm whale groups consisting of females, calves and immature animals of both sexes are encountered in sub-tropical and tropical waters. Codas are generally heard from these groups during periods of apparently social behaviour at or near the surface, behaviour that contrasts sharply with the prolonged dives and wide spacing of foraging groups (Whitehead and Weilgart 1991).

Only a handful of studies have been made of these vocalisations to date (Moore et al. 1993; Watkins & Schevill 1977; Weilgart & Whitehead 1993, 1997; Whitehead et al. 1998), and none evaluated the analytical methods they used. Initially codas were assigned to classes by simple observation and judgement (e.g. Moore et al. 1993; Watkins & Schevill 1977); the underlying assumption that the classes were real and meaningful to the animals themselves was suggested by the extreme stereotypy of the coda patterns. More recently, Weilgart & Whitehead (1997) used *k*-means cluster analysis. Both these methods come with pitfalls. The human 'eyeball' method contains two assumptions: that what seems different to us is actually different to the animals, and that what seems different to one person will also seem different to another observer. The former is rarely tested in animal bioacoustics and certainly has not been for sperm whales, while the latter is testable (Janik 1999) and must be met if the essential scientific criterion of repeatability is to be fulfilled. The *k*-means cluster analysis used by Weilgart & Whitehead (1997), for all its numerical objectivity, comes with the problem of determining k – the number of clusters into which the data are to be grouped. Weilgart & Whitehead (1997) used a fixed number of clusters (5 for 3-click codas and 10 for >3 click codas) and then lumped all clusters with less than 50 codas into a catch-all 'variable' category. They then compared numbers of codas in each class between different social groups. While objective, this methodology obviously discards potentially interesting information in the form of rarer coda classes.

Both classification-based methods, while making data easier to understand given our aptitude for categorisation (Tomasello 1999, pp.17-18), carry the underlying assumption that real 'types' are present. However, this is not necessarily the case for other species. In cetaceans, for example, Murray et al. (1998) showed that the calls of false killer whales *Pseudorca crassidens* form a graded sequence with no clear divisions. Similarly, pilot whale *Globicephala melas* whistles appear to form a graded continuum between several basic types (Taruski 1979). We can use empirical cues to justify a decision to classify – for example if calls are stereotyped with few or no intermediate forms. However, if methods of comparing sets of vocalisations that do not rely on classification are available then one can employ both classification and non-classification approaches in

tandem for a rigorous investigation; conclusions supported by analyses using both approaches are concomitantly stronger. Here we explore methods of classifying codas and of comparing repertoires using classification as well as non-categorical methods.

METHODS

Data collection

In this study we used a subset of codas recorded from field studies around the Galápagos Islands. For general field methodology see Whitehead & Weilgart (2000). Codas were recorded using one of two sets of equipment. The first was an Offshore Acoustics hydrophone (frequency response: 6 Hz -10 kHz ±3 dB) connected to a Sony TC-D5M cassette recorder, used for the 1999 recordings of social unit "T"; the second consisted of a Benthos AQ17 hydrophone (1-10 kHz), connected via either Barcus-Berry 'Standard' or Ithaca 453 pre-amplifiers to either a Uher 4000, Sony TC770 or Nagra IV-SJ recorder, used for the 1985 and 1987 recordings of social units "A" and "B". Recordings were digitised at 44.1 kHz onto a standard desktop PC, and we analysed codas using a software package called Rainbow Click (Gillespie 1997; Leaper et al. 2000) specifically developed for the study of sperm whale sounds (e.g. Jaquet et al. 2001). The software detects clicks using a two level trigger with user-variable parameters and then stores the detected clicks in a data file. The timing of clicks within codas can then be extracted once codas have been defined and marked individually by the user. Only codas that could be unambiguously heard (at the various playback speeds supported by the software) were marked, so some codas that were recorded were not analysed due to a variety of factors leading to a generally poor recording quality (these included water noise, engine noise and overlapping by other clicks and codas). The resultant data for each coda were the absolute inter-click intervals, defined as the time between the onsets of consecutive clicks, so for example a four click coda that we analysed was stored as 0.180, 0.178, 0.182 (units are seconds). These data were then standardised to coda length by dividing each interval by the total length of the coda (defined as the time between the onsets of the first and last clicks). This was done because previous work has shown coda rhythm to be better preserved than tempo (Moore et al. 1993) and so most work on codas discards tempo information (e.g. Weilgart & Whitehead 1997). It is therefore an assumption of this paper and the methods we describe that it is the rhythm of clicks within a coda and not the tempo that is biologically important and thus of interest. For the present analysis we used a sample of 1548 codas from our database of analysed codas (Table 1) that were assigned to social units based on the

presence of photographically identified individual whales (Christal et al. 1998).

Observer classification

Janik (1999, 2000) has shown that human classification, with all its pitfalls of arbitrariness, is still the best way to classify bottlenose dolphin (*Tursiops* sp.) signature whistle contours. We therefore emulated his methods by using three people (one of us – LER – and two volunteers) to independently classify codas. Each observer was presented with a computer display of the coda to be classified (on a standardised scale so that tempo information was discarded for this method as well) and assigned codas classes as they saw fit based on their perception of the classes present in the dataset. There was no limit on the number of classes, and at any point observers could view the mean of any already existing class as well as a display of the current coda alongside all the other codas with the same number of clicks in the analysis set. For this method we used only the 879 codas from social unit T, in order to keep the task manageable. Once all three independent classifications were complete, the results were scanned for common classes and if two or more observers agreed on a class for a given coda then it was assigned to that class, while if there was no agreement then the coda was dropped from further analysis. If significant proportions of codas are rejected on this basis then it becomes clear that this methodology is not as applicable to sperm whale codas as to bottlenose dolphin whistles. Such levels of rejection may also suggest that perhaps coda types are not as discrete as once thought.

TABLE 1

Data used in this study. Social unit codes correspond to those in Christal et al. (1998) and Christal & Whitehead (2001)

Social Unit Code	Number of recordings	Dates recorded (first – last)	Number of codas
A	25	24 February 1985 – 9 March 1987	572
B	9	23 January 1987 – 22 March 1987	97
T	22	10 March 1999 – 10 April 1999	879
		Total:	1548

K-means clustering

Using automatic classification algorithms avoids the problems of subjectivity inherent when humans classify codas. Weilgart & Whitehead (1997) used k-means clustering, where data are divided into k clusters so as to minimise the pooled within-cluster sum of squares. Such analysis has to treat codas with differing numbers of clicks separately – i.e. four-click codas will be clustered in a separate analysis from five-click codas. This is because four- and five-click codas represent multivariate datasets with different numbers of dimensions, in this case three and four dimensions, since an n click coda can be represented by n-1 standardised click intervals. While codas can theoretically be represented by n-2 standardised intervals (since all intervals must sum to one) this biases distance measures to emphasise differences that occur in the first n-2 intervals over differences in the last interval, so we included all n-1 intervals in our analyses. However, the problem of deciding k non-arbitrarily remains. Weilgart & Whitehead (1997), as noted above, adopted a technique that involved discarding some potentially important information. Schreer et al. (1998) attempted to use a 'stopping rule' based on Calinski & Harabasz's (1974) Variance Ratio Criterion (VRC). Although they found it unsatisfactory for their data, we too tried this stopping rule based on the VRC:

$$VRC = \frac{BGSS}{k-1} \Big/ \frac{WGSS}{n-k} \tag{1}$$

where *BGSS* and *WGSS* are the between and within group sum of squares respectively, k is the number of clusters and n is the number of observations. We ran an iterative k-means clustering algorithm on each coda size (4 click, 5 click etc) for $2 \leq k \leq 10$. In this and all the k-means analyses that we ran, initial cluster centroids were selected at random from the input data. Since the iterative k-means algorithm does not necessarily always converge on the optimal solution, each clustering was run 10 times and the solution with the lowest *WGSS* selected and retained. We then calculated the VRC for each solution. Calinski & Harabasz (1974) suggest that the optimal clustering solution is at the first local maximum of the VRC as k increases. However, as we shall show later, we encountered the same problem as Schreer et al. (1998): the VRC did not give clear, unambiguous results for our test data. In Milligan & Cooper's (1985) comparison of stopping rules, the VRC rule performed best at detecting the number of clusters in sample datasets. However, their sample data were very strongly clustered, and these authors openly attempted to let every stopping

rule 'adopt the most favourable conditions' to optimise its performance while cautioning that their 'findings are likely to be somewhat data dependent'. Hence, we also tried the rule that performed next best after the VRC: Duda and Hart's (1973; pp.239-243) ratio criterion. This criterion tests the null hypothesis that the partitioning of a given dataset into two clusters is spurious at the p-percent level, and rejects that null hypothesis if

$$\frac{WGSS_{(2)}}{TSS} < 1 - \frac{2}{\pi m} - \alpha\sqrt{\frac{2\left(1 - 8/\pi^2 m\right)}{nm}} \qquad (2)$$

where TSS is the summed squared deviation from the mean for the unclustered data, $WGSS_{(2)}$ is the pooled within-cluster sum of squares for the same data in two clusters ($J_e(1)$ and $J_e(2)$ respectively in Duda & Hart's (1973) notation), m is the number of dimensions in the data, n is the number of observations and α is a standard normal score given by

$$p = 100\int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-1/2u^2} \delta u \qquad (3)$$

This measure compares the reduction in the squared error, as given by the ratio $WGSS_{(2)}/TSS$, against the distribution of reductions expected from dividing a multivariate normal population through the mean. Note that the method only makes decisions about dividing a given set into two clusters. This gives it one major heuristic advantage over the VRC method in that it provides a basis for deciding whether any clustering at all is justified, i.e. the first split of the original data into two clusters. We used this criterion in a divisive procedure, in contrast to the VRC method, which seeks globally optimal solutions for the entire dataset. Data were repeatedly split using iterative k-means with k = 2 (repeated 10 times, selecting the lowest $WGSS$ solution, as above). Each split was then accepted or rejected with p = 95%, and the resultant clusters again split and tested. Division continued until no cluster could be split according to the $WGSS_{(2)}/TSS$ criterion.

Once we had arrived at two classifications based on human and divisive k-means methods (the latter were performed on the expanded dataset of 1548 codas), we compared the human classification (of codas on which at least two observers had agreed) with the k-means classification of those same codas using Cramer's V (Wilkinson et al. 1996). While this metric does not provide for any kind of significance testing, it does give a relative measure of how well two classifications coincide. We also calculated Cramer's V for each individual human

against each other, against the 'consensus' human classification and against the k-means classification.

Classification-free approach

One obvious way to compare codas without resorting to classification is by using distances between codas in multivariate space. We tested two different ways of measuring distances between vectors representing points in multivariate space: Euclidean distance and the infinity-norm. The Euclidean distance (dE_{ij}) between codas i and j is defined as

$$dE_{ij} = \sqrt{\sum_{k=1}^{c} \left(x_{ik} - x_{jk}\right)^2} \tag{4}$$

where c is the number of standardised click intervals representing codas i and j (i.e. the number of clicks minus one), x_{ik} is the k^{th} interval of coda i and x_{jk} is the k^{th} interval of coda j. The infinity-norm distance (dI_{ij}) is defined as the maximum absolute difference between the vectors x_i and x_j (sometimes written as $\|x_i - x_j\|_\infty$). Both these metrics are direct measures of how dissimilar codas i and j are (i.e. low values mean that codas i and j are nearly identical in pattern). However, both can in theory lead to results that appear counter to our stated aim of comparing coda rhythms. In the case of Euclidean distance, consider a regular five-click coda (5R): perturbing all the clicks by some small amount (x) results in a slightly irregular coda that still has a generally regular rhythm, while perturbing a single click by a large amount (y) gives a distinctly different rhythm (such as 4+1). However, the codas resulting from these perturbations could have very similar Euclidean distances from the original if $x \approx y/\sqrt{(n-1)}$, where n is the number of clicks (in this case five), despite having quite different rhythms. In the case of the infinity-norm distance, consider again a 5R coda, again perturbing one of the clicks by some amount (y) to produce, for example, a 4+1 rhythm. Then consider perturbing two of the original five regular clicks by a smaller amount (x) to produce, for example, a 3+1+1 rhythm. If $y > x$ then the 3+1+1 coda will have a smaller distance from the original 5R than the 4+1, even though a 3+1+1 rhythm is arguably less similar to 5R than is a 4+1 rhythm. Thus neither metric always directly quantifies rhythmic differences in a consistent way. Whether these situations are occurring enough to significantly impact on results depends on the actual codas. One way to test whether these theoretical problems have a significant impact on results is to use both on the same data – if the two techniques produce

similar patterns of results, then it is unlikely that the theoretical conditions outlined above are occurring much in practice.

The above metrics reflect the differences between pairs of codas, but we are interested in comparing sets, or repertoires, of codas. One measure of how dissimilar any one repertoire of codas is from another is the mean distance defined as the average of pairwise distances between the two repertoires. That is, for each coda in repertoire A, calculate the distance to all other codas in repertoire B, and take the mean of all the resultant values (size of repertoire A x size of repertoire B), or more formally,

$$S_{AB} = \frac{\sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d_{ij}}{n_A . n_B} \tag{5}$$

where n_A and n_B are the number of codas in repertoire A and B respectively and d_{ij} can be either the Euclidean or infinity-norm distance. However, we cannot use this directly, because repertoires contain codas of differing sizes, that is, codas with *different* numbers of clicks – it is thus impossible to measure a direct multivariate distance between them. One could overcome this by simply taking the mean of all the distances between codas of the same size, but this would not take into the account the differences in numbers of codas of different sizes between the repertoires. Using distances, comparisons between codas with different sizes could be set to an arbitrarily high number, but then average distances would depend more on the arbitrary value of this number than any other factor. Alternatively, one can use similarity scores that are inversely proportional to distance; for example, $b/(b + d_{ij})$ is a measure of similarity, where the value of b relative to the spread of data gives the approximate resolution at which the measure operates. If comparisons are expressed as similarities rather than distances, then comparisons between codas of different sizes can be simply set to zero. We took this approach, rendering every comparison between codas of different sizes zero, and then took the mean of all comparisons (not only those between codas of the same size). Equation (5) thus becomes

$$S_{AB} = \frac{\sum_{i=1}^{n_A} \sum_{\substack{j=1 \\ l_j = l_i}}^{n_B} \frac{b}{b + d_{ij}}}{n_A . n_B} \tag{6}$$

where l_i is the number of clicks in coda i of repertoire A and l_j is the number of clicks in coda j of repertoire B.

Using this approach, a repertoire can also be compared with itself; if A=B then equation (6) gives S_{AA}, the self-similarity. This is important because, unlike most similarity measures, the results of comparing a repertoire with itself using equation (6) are not readily predictable. Equation (6) does not produce 1 when repertoires are compared with themselves, instead it gives an approximate indication of the 'spread' or diversity of a given repertoire; relatively compact repertoires will have relatively high self-similarities. For this work however, the clear implication is that values of S_{AB} for between-repertoire comparisons should be interpreted alongside the values of S_{AA} and S_{BB} calculated when those repertoires are compared to themselves, unless large numbers of comparisons are being made in which case it would be more tractable to enter the similarity measures into a hierarchical cluster analysis.

We calculated comparisons between repertoires of codas recorded from different social units so as to compare the results from this technique with results from correlating classified codas as described in the previous section. We used equation (6) with both distance metrics, and b = 0.001, 0.01, 0.1 and 1, to look at how the different metrics and different values of b change the similarity results. We also compared repertoires using the results of the classification methods by calculating Spearman rank correlation coefficients on counts of how many codas of each type were heard in each repertoire (as in Weilgart & Whitehead 1997). When comparing repertoires between social units using both similarity and classification methods, we estimated the robustness of each measure by calculating bootstrap standard errors from 100 random samples with replacement (Sokal & Rohlf 1995). All the numerical procedures described here were implemented in MATLAB (v12.0; we encourage interested researchers to contact us for copies of the MATLAB routines), with the exception of Cramer's V which was calculated in SYSTAT (v10), on a standard PC.

RESULTS

Observer and *k*-means classification

After classification by three people, 861 of 879 codas (98.0%) from social unit T met the consensus criteria of agreement by at least two observers, and were classified into 49 types. 85% of the codas recorded had 6 or less clicks, and while the 8 most common types account for 82.1% of the repertoire (the single most common type accounting for 19.9%, approximately 1/5 of all the codas heard), there were many rare

types – 36 of the 49 types individually made up less than 1% of the codas recorded.

As we mentioned above, the VRC stopping rule did not give unambiguous results (Figure 1); VRC values for various k did not show unambiguous local maxima as described in Calinski & Harabasz (1974). The divisive procedure, which runs unsupervised and does not require any input apart from the initial acceptance threshold (p) for the Duda and Hart criterion, produced 32 clusters from the entire 1548 coda dataset. Here the most common type accounted for 11% of codas recorded, the 8 most common types for 45%, and 21 types individually made up less than 1% of the codas recorded. Generally, the divisive method produced clusters that match reasonably with observable clumping of the data (Figure 2).

Cramer's V statistics for each combination of classifications are given in Table 2. It is noteworthy that all three human observers individually produced results more similar to the k-means procedure than to any of the other observers. To illustrate how this relates to the actual data, Figure 3 compares the classification of codas with four clicks by the k-means and observer consensus methods. Note both the

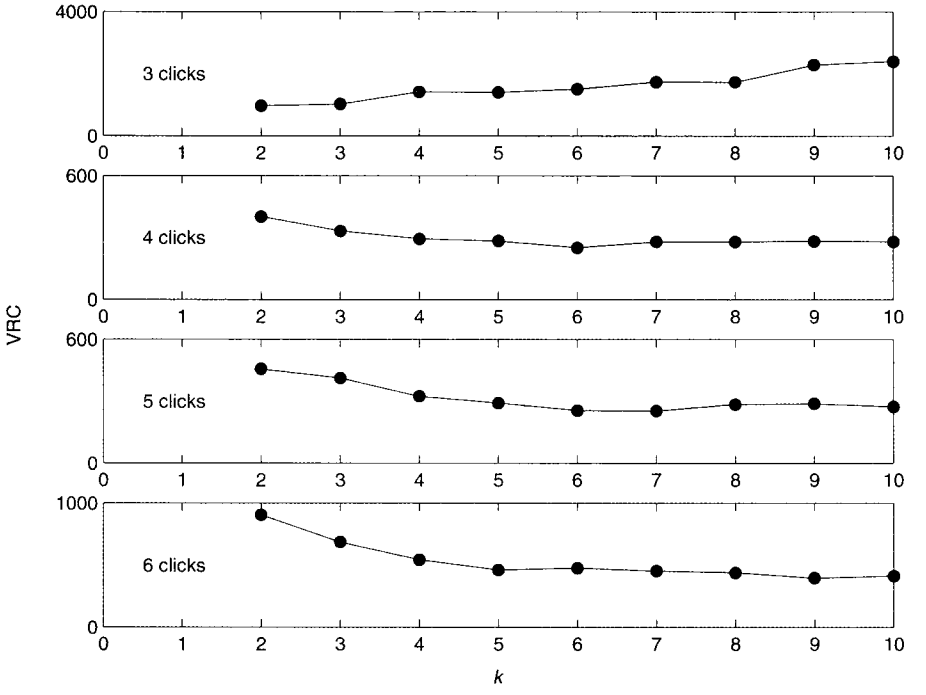


Figure 1. Variance Ratio Criterion values from k-means solutions for $2 \leq k \leq 10$, calculated using codas with 3-6 clicks. Note the lack of clear local maxima.

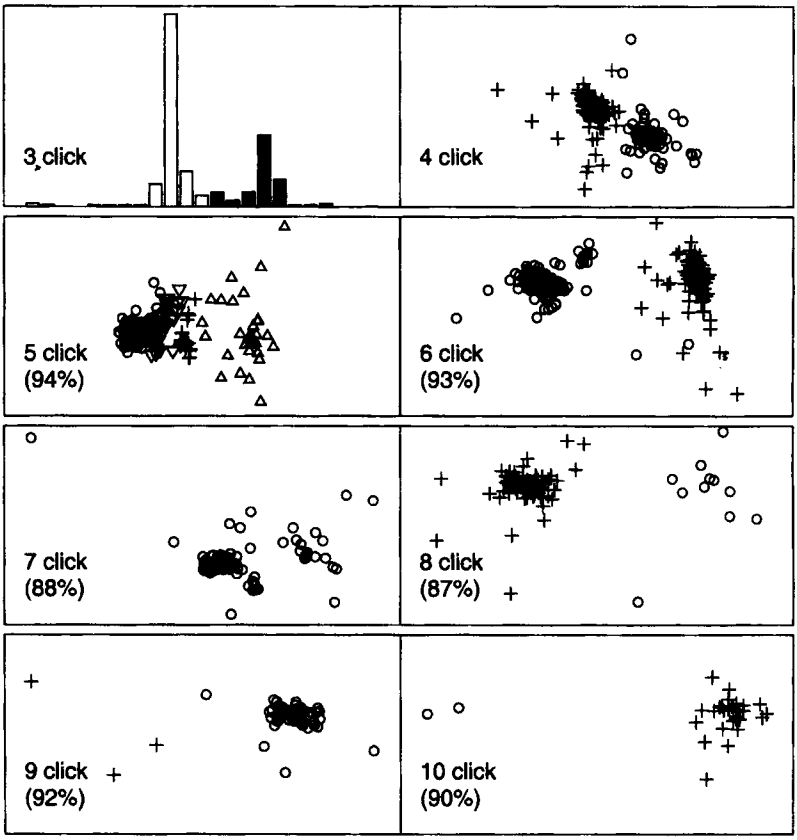


Figure 2. Results of divisive k-means clustering using the Duda and Hart criterion. For 3 click codas, the plot shows frequency distribution of the first standardised click interval (SCI), with different clusters having different shaded bars. For 4 click codas, the first SCI is plotted against the second SCI. All the other plots show the first two principal components derived from SCIs, along with the percent variance accounted for by those first two principal components. Different clusters are represented by different symbols.

TABLE 2

Cramer's V, for each combination of classifications. Observer A, B and C are individual classifications, Human-All is the consensus and K-means the results of the divisive k-means procedure. Higher values represent a relatively higher correspondence between classifications.

	K-means	Human – All	Human – A	Human – B
K-means				
Human – All	0.95			
Human – A	0.95	0.95		
Human – B	0.93	0.96	0.91	
Human – C	0.96	0.91	0.88	0.87

clear structuring of the data in the plot, and also how the '4R' type in the observer classification considerably overlaps the '3+1' cluster, while the *k*-means results produce (not surprisingly) relatively well-separated clusters.

Multivariate similarity

Figures 4a and 4b show the results of employing the similarity measures to compare repertoires between social units, using Euclidean and infinity-norm distances respectively, and for a range of *b*. The results are plotted along with the self-similarity scores for each repertoire. We would consider any result where the comparison value

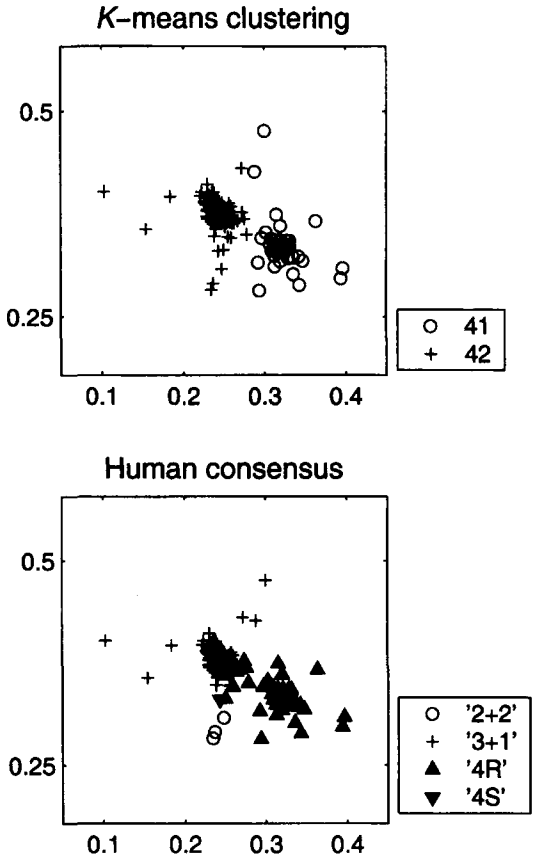


Figure 3. First SCI plotted against the second SCI for all four click codas from social unit T, plotted by cluster membership as determined by divisive *k*-means or human consensus classification.

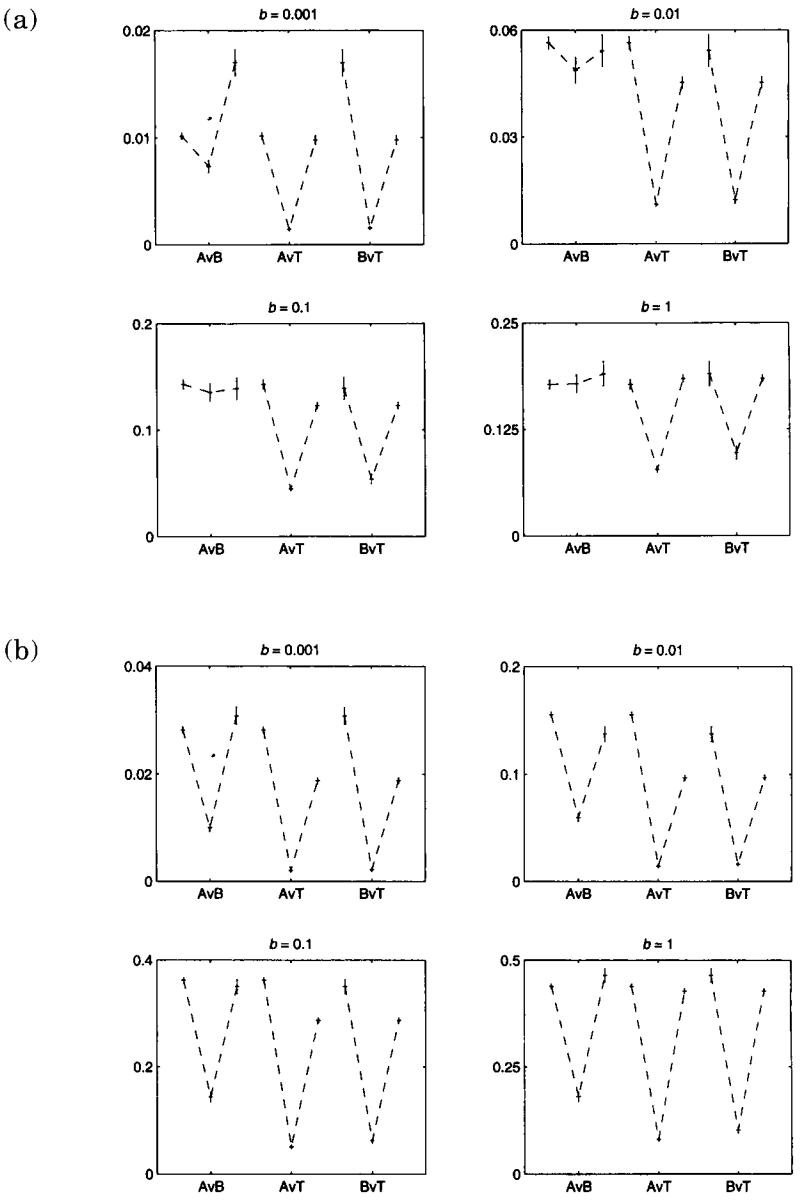


Figure 4. Repertoire similarities calculated between social units for various values of b. Each dotted line joins the comparison similarity to self-similarity values for both social units in the comparison – for example the leftmost line in each plot joins the self-similarity of unit A, the comparison similarity between A and B and the self-similarity of B. Error bars are standard errors from 100 bootstrap samples.

(a) Similarities calculated using the Euclidean distance.

(b) Similarities calculated using the infinity-norm distance.

lay in the bootstrap standard error range of the two self-similarity values to indicate that the two repertoires were statistically indistinguishable (at the resolution set by the value of b). The pattern of results is identical for both distance metrics and for all values of b: the repertoires of social units A and B are more similar to each other than either are to unit T, although when using Euclidean distance with higher values of b the repertoires of A and B are indistinguishable. This pattern agrees well with comparisons using correlations between repertoires of codas classified using the divisive k-means method, where the correlation between A and B are higher than either with T (Figure 5). To show how these results reflect the true nature of the underlying data, Figures 6a and 6b show plots of three to ten-click codas for social units A and B, and A and T respectively. It is clear from these plots that the repertoires of A and B overlap each other considerably. In contrast, while there is some overlap between A and T, there are also clear areas of non-overlap, particularly with four, five, and six click codas.

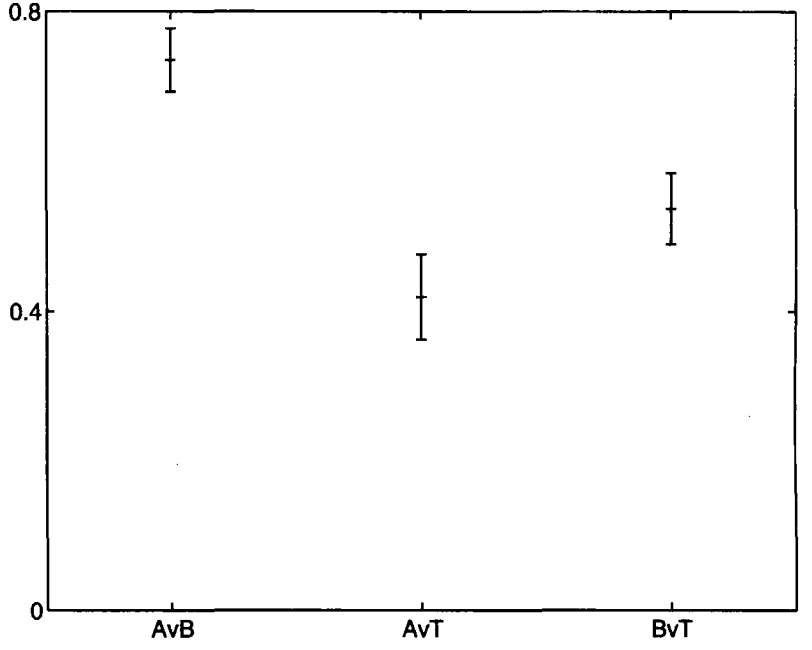
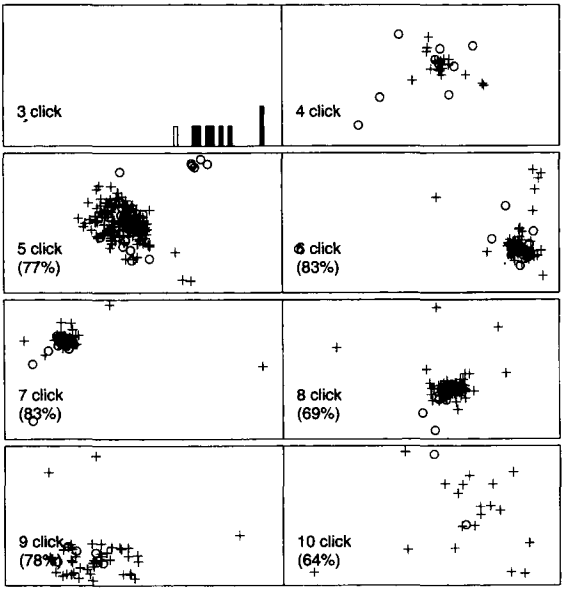


Figure 5. Spearman correlation coefficients for comparisons between repertoires of codas classified by divisive k-means, calculated as in Weilgart & Whitehead (1997). Error bars show standard errors from 100 bootstrap samples.

(a)



(b)

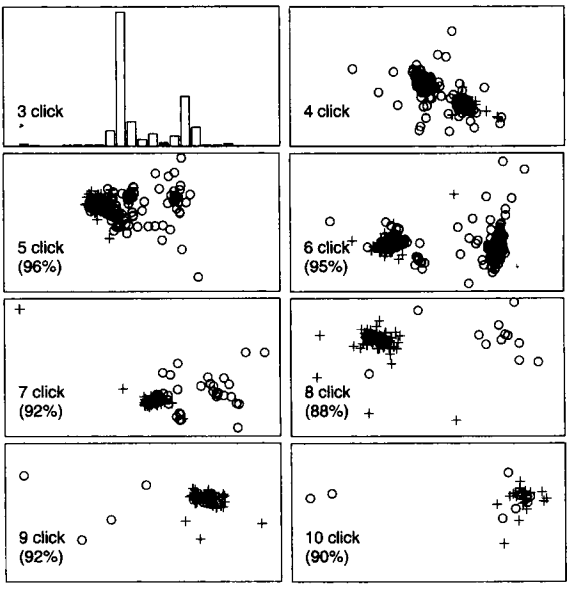


Figure 6. Plots comparing coda repertoires between social units. For 3 click codas, the plot shows frequency distribution of the first standardised click interval (SCI). For 4 click codas, first SCI is plotted against the second SCI. All the other plots show the first two principal components derived from SCIs, along with the percent variance accounted for by those first two principal components.

(a) Unit A (+, shaded bars) and unit B (o, clear bars).

(b) Unit A (+, shaded bars) and unit T (o, clear bars).

DISCUSSION

Of the methods that we implemented or attempted to implement (observer classification, VRC k-means, divisive k-means using the Duda and Hart criterion, and multivariate similarity), only the latter three are likely to be useful in future studies of sperm whale codas. While observer classification has been shown to work rather well for other cetacean vocal studies (Janik 1999, 2000), it is apparently not so useful for studying sperm whale codas, for several reasons. Firstly and most importantly, humans did not pick out the naturally occurring groupings in the data as well as divisive k-means method. For example, the human defined classes stretch across the two main clusters evident in the four-click codas (Figure 3). Secondly, the Cramer's V results for the classifications show that the human classifications were inconsistent with respect to each other, suggesting that the repeatability of these measures would not be especially robust. If the acceptance criteria were raised to require agreement from all three observers, only 34% of the codas would be accepted, which casts further doubt on the robustness of the technique. We can only speculate as to why this might be so, but one possible reason might be the large amounts of data used here: we asked humans to classify 879 codas, while Janik (1999) used only 104 bottlenose dolphin (*Tursiops* sp.) signature whistles in his study. Remembering one's previous classifications is likely much easier for smaller datasets, particularly since Janik (1999) also printed hard copies of each whistle spectrogram for the comparison exercise, something which was unfeasible for the 879 codas we used here and more so for the larger datasets we would like to use these methods on in future. Finally, the observer classification method also involved the rejection of an albeit small number of codas; we cannot justify throwing away information in this way given the difficulty and expense of making these recordings in the first place, nor given the possibility of introducing bias if classifiers are more likely to disagree over certain forms of coda than others. Hence we do not see observer classification as a useful method in this particular case.

We do think that some form of classification is justified, given the structure present in the data (Figures 2, 3 and 6) – there do seem to be some very tightly defined coda 'types'. Both the k-means methods potentially provide a robust method for classifying codas that will produce repeatable results. However, the VRC stopping rule did not seem to work well here: while it performs excellently with well-defined clusters (Milligan & Cooper 1985), we found that the rule produced ambiguous results for our test data. In contrast, the divisive k-means method assigned data to clusters that matched the structuring of the dataset quite well (Figure 2). In addition, comparisons between the three social units based on this classification produced results that

make sense with respect to the underlying data (Figure 6). While this latter technique is clearly better in this case, we would add the caveat that it might not always be better. As Milligan & Cooper (1985) point out, the performance of any such criterion is very likely to be data dependent, and while it appears that for the present data the VRC rule is not the most appropriate, this may not always be so. For example, a dataset with several clear clusters may not produce a significant difference in Duda and Hart's ratio criterion on the first split into just two groups, and thus the split may be rejected even though clustering is obviously present to a human observer. We therefore suggest that the results of both techniques be checked against the raw data to ensure that logical clusters are being retrieved. The ultimate choice of technique will be data dependent and somewhat arbitrary, based on an observer's judgement of how well the clustering solution fits the data. One disadvantage with these clustering methods is that care must be taken with the *ad hoc* addition of new data. One could classify new data using Mahalanobis or Euclidean distances to assign new codas to the cluster with the nearest centroid, but only for small amounts of new data. Visual inspection (e.g. as in Figure 2) would be necessary to ensure that new, very different, codas were not being 'forced' into existing categories, and the entire procedure should be run again if large amounts of new data, or data very different from that used for the original clustering, are added. While there are many other clustering algorithms available, as well as more recent developments in artificial neural networks (see e.g. Deecke et al. 2000) we leave it to other interested researchers to investigate their viability in this application; the simplicity and wide recognition of k-means, along with the reasonable results given here, make it suitable for our purposes. In the only such study of which we are aware on biological data, Schreer et al. (1998) concluded that k-means was the best classification technique for dive profile data in an analysis that included performance comparisons with artificial neural networks, although obviously there are major differences between dive profile and coda data.

It is encouraging that the multivariate similarity measures we devised show the same pattern as the classification methods – this agreement gives a greater confidence that the results are robust, particularly since the pattern of results is repeated across all values of b. It is also encouraging that the similarity results comparing social unit repertoires reflect rather well the degrees of overlap evident in Figure 6. The two distance metrics that we tested produced very similar patterns of results, with the main difference being that the infinity-norm distance was perhaps the better discriminator across all values of b, and so may give a more precise measure of repertoire similarity. We also argue that the similar patterns of results suggest that the theoretical limitations of both distance metrics are not

substantially affecting results. While the possibility that unforeseen combinations of codas may produce results dissonant with our stated aim of comparing coda rhythms still exists, we have shown that the methods produce results consistent with a different analysis method, *k*-means clustering (Figures 4 & 5), and with observable patterns in the raw data (Figure 6); the aggregative nature of our measure likely leads to specific anomalies being subsumed in broader scale patterns. One drawback of this method is that it is computationally demanding, especially for high numbers of bootstrap resamples: we performed the analyses again with 1,000 bootstrap resamples (as opposed to the original 100), which took our computer approximately three days to complete. The bootstrap standard errors for 1,000 resamples were nearly identical to those for 100 samples, but sometimes smaller.

The slightly different results produced by various values of *b* are interesting: with similarity defined as in Equation 6, the value of *b* is an approximate measure of the *resolution* at which comparisons are being made, in terms of normalised inter-click interval. Considering Figure 3, and noting that the two obvious clusters present are about 0.1 standardised click-interval units apart, suggests that calculating similarities with $b = 0.1$ will likely give us information on whether the

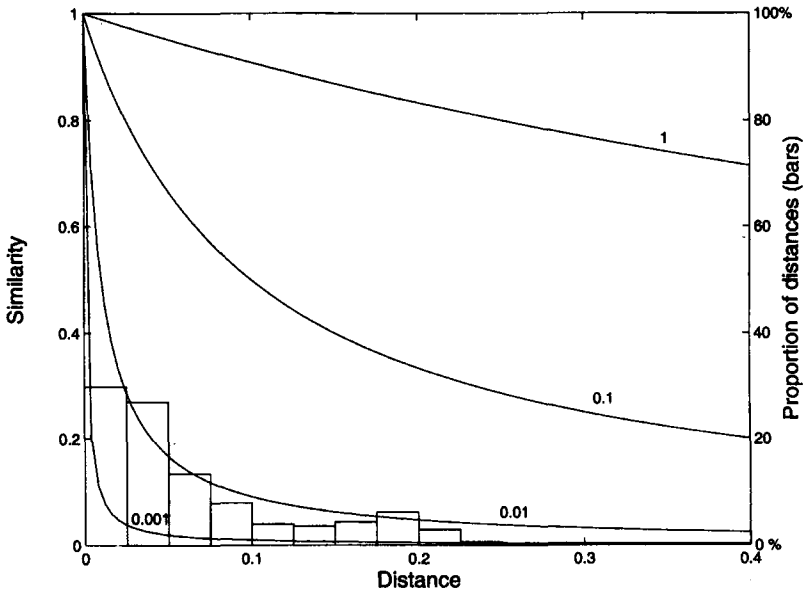


Figure 7. Plot of the functional form of the similarity transformation at various *b* values (labelled lines; distance on the x-axis and resultant similarity on the y axis) overlaid with a histogram of distances between the five click codas on our dataset calculated using the infinity-norm (data are proportion of total).

two social units make codas in similar or different clusters. Decreasing b to 0.01 or even 0.001 gives a very fine scale comparison of exactly how well the codas in each repertoire coincide within clusters, since each cluster in Figure 3 is approximately 0.05 data units across. Plotting the functional form of the similarity transformation against an actual distribution of coda distances shows how the transformation emphasis shifts to smaller distances as b decreases (Figure 7). It therefore does not make sense to recommend a fixed value for b as different values can provide information at different scales of analysis. In our data actual coda lengths, defined as the time between the onsets of the first and last clicks, ranged from 0.189 s to 9.510 s, with a mean of 1.228 s. So $b = 0.001$ corresponds on average to a resolution of 1.2 ms (range 0.2-9 ms), which roughly equals the maximum resolution of our analysis system. Whether sperm whales can detect rhythmic differences of this scale remains a moot question.

During this work we also developed and tested another approach to measuring similarities between codas of different sizes. This approach arose from the observation that certain coda 'classes' identified by Weilgart & Whitehead (1997) seem to span various coda sizes. For example, the '+1' codas (click-click-click-pause-click would be a 3+1 coda) are heard with differing numbers of clicks (i.e. 3+1, 4+1, 5+1). To reflect such classes in our similarity measure we 'cross-correlated' codas of different sizes using a technique best explained by example. Consider two codas, A and B; A contains four clicks and B contains six. Without disturbing the order or neighbour-relationships of clicks in B, one can extract 3 different four-click codas from it: B_1 with clicks one to four, B_2 with clicks two to five and B_3 with clicks three to six. Each of these can be standardised by dividing each click interval by the sum of the absolute click intervals of each given subset (B_1, B_2, B_3). One can then 'cross-correlate' the two codas by calculating the Euclidean distance between A and (B_1, B_2, B_3) and taking the minimum distance (hence maximum similarity). This similarity was then 'discounted' by an amount related to the difference in number of clicks between codas (otherwise, for example, a 3R and a 15R coda could have equal similarity to a 3R and another 3R, which few would consider useful); so instead of having similarities between codas of different size rendered zero as in Equation 6, a discounted distance would be entered into the averaging. Thus a 3+1 coda would be scored more similar to a 4+1 coda than to a 5R coda. However, reviewers pointed out that this method is biased toward uniform rhythm patterns, and does not always give results consistent with the aim of comparing rhythm. For example, consider a 4+1 and a 5R coda compared to a 4R coda: the 4+1 and 5R coda could produce identical similarities to the 4R by using only the first four clicks of each, despite having quite different rhythms. This theoretical inconsistency together with a more than three-fold increase in computation time to produce

results that had the same pattern as Figure 4 led to us deciding that this was unlikely to be a useful method in the future.

In conclusion, we have developed and/or tested potentially useful ways to compare collections, or repertoires, of codas. The methods can be used in a variety of ways such as comparing social unit repertoires as here, or to compare the coda output of the same group recorded at different times, or in different ecological or behavioural situations. In addition, the methods can be employed at levels both below and above that of the group – from individual repertoires (if codas can be reliably assigned to individual whales) through to comparisons between oceans. In future studies it would be advisable to use both categorical and non-categorical techniques in tandem in order to minimise concerns that results are simply due to spurious categorisations or weaknesses in the distance metrics. We hope that these techniques will be helpful in future studies of sperm whale codas and other studies that face similar problems in comparing vocal repertoires represented by multivariate datasets.

ACKNOWLEDGEMENTS

We are extremely grateful to the International Fund for Animal Welfare for allowing us to use Rainbow Click and in particular Doug Gillespie for helping us by modifying the software to meet our needs, and to Amanda Coakes and Tonya Wimmer for classifying codas. We also thank Steve Henderson for mathematical advice. Amanda Coakes, Meaghan Jankowski, Kurt Fristrup and four anonymous reviewers (one of whom was especially helpful in suggesting the infinity-norm distance measure) gave useful comments on the manuscript. The Natural Sciences and Engineering Research Council of Canada supported the fieldwork and LER was supported by a Canadian Commonwealth Scholarship and an Izaak Walton Killam Memorial Scholarship.

REFERENCES

Calinski, T. & Harabasz, J. (1974). A dendrite method for cluster analysis. *Comm. Statist.*, **3**, 1-27.

Christal, J. & Whitehead, H. (2001). Social affiliations within sperm whale (*Physeter macrocephalus*) groups. *Ethology*, **107**, 323-340.

Christal, J., Whitehead, H. & Lettevall, E. (1998). Sperm whale social units: variation and change. *Can. J. Zool.*, **76**, 1431-1440.

Deecke, V. B., Ford, J. K. B. & Spong, P. (2000). Dialect change in resident killer whales: Implications for vocal learning and cultural transmission. *Anim. Behav.*, **40**, 629-638.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York: John Wiley & Sons.

Gillespie, D. (1997). An acoustic survey for sperm whales in the Southern Ocean sanctuary conducted from the *R/V Aurora Australis*. *Rep. int. Whal. Commn*, **47**, 897-908.

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. (2000). Whistle matching in wild bottlenose dolphins (*Tursiops truncatus*). *Science*, **289**, 1355-1357.

Jaquet, N., Dawson, S. & Douglas, L. (2001). Vocal behavior of male sperm whales: Why do they click? *J. Acoust. Soc. Am.*, **109**, 2254-2259.

Leaper, R., Gillespie, D. & Papastavrou, V. (2000). Results of passive acoustic surveys for odontocetes in the Southern Ocean. *J. Cetacean Res. Manage.*, **2**, 187-196.

Milligan, G. W. & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159-179.

Moore, K. E., Watkins, W. A. & Tyack, P. L. (1993). Pattern similarity in shared codas from sperm whales (*Physeter catodon*). *Mar. Mammal Sci.*, **9**, 1-9.

Murray, S. O., Mercado, E. & Roitblat, H. L. (1998). Characterizing the graded structure of false killer whale (*Psuedorca crassidens*) vocalizations. *J. Acoust. Soc. Am.*, **104**, 1679-1687.

Nowicki, S. & Nelson, D. A. (1990). Defining natural categories in acoustic signals: Comparison of three methods applied to 'Chick-a-dee' calls. *Ethology*, **86**, 89-101.

Schreer, J. F., O'Hara Hines, R. J. & Kovacs, K. M. (1998). Classification of dive profiles: A comparison of statistical clustering techniques and unsupervised artificial neural networks. *J. Agric. Biol. Envir. S.*, **3**, 383-404.

Sokal, R. R. & Rohlf, F. J. (1995). *Biometry*, 3rd ed. San Francisco: W. H. Freeman.

Taruski, A. G. (1979). The whistle repertoire of the North Atlantic Pilot Whale (*Globicephala melaena*) and its relationship to behaviour and environment. In *Behaviour of Marine Mammals* (Ed. by H. E. Winn & B. L. Olla). New York: Plenum Press.

Terhune, J. M., Burton, H. & Green, K. (1993). Classification of diverse call types using cluster analysis techniques. *Bioacoustics*, **4**, 245-258.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge, Mass.: Harvard University Press.

Watkins, W. A. & Schevill, W. E. (1977). Sperm whale codas. *J. Acoust. Soc. Am.*, **62**, 1486-1490.

Weilgart, L. & Whitehead, H. (1993). Coda vocalizations in sperm whales (*Physeter macrocephalus*) off the Galapagos Islands. *Can. J. Zool.*, **71**, 744-752.

Weilgart, L. & Whitehead, H. (1997). Group-specific dialects and geographical variation in coda repertoire in South Pacific sperm whales. *Behav. Ecol. Sociobiol.*, **40**, 277-285.

Whitehead, H., Dillon, M., Dufault, S., Weilgart, L. & Wright, J. (1998). Non-geographically based population structure of South Pacific sperm whales: dialects, fluke-markings and genetics. *J. Anim. Ecol.*, **67**, 253-262.

Whitehead, H. & Weilgart, L. (1991). Patterns of visually observable behaviour and vocalizations in groups of female sperm whales. *Behaviour*, **118**, 275-296.

Whitehead, H. & Weilgart, L. (2000). The sperm whale: Social females and roving males. In *Cetacean Societies* (Ed. by J. Mann, R. C. Connor, P. Tyack & H. Whitehead). Chicago: University of Chicago Press.

Wilkinson, L., Blank, G. & Gruber, C. (1996). *Desktop Data Analysis with SYSTAT*. New Jersey: Prentice Hall.

Received 17 January 2002, revised 17 June 2002 and accepted 13 December 2002