ANIMAL BEHAVIOUR, 1999, 57, 133–143
Article No. anbe.1998.0923, available online at http://www.idealibrary.com on IDEAL

Pitfalls in the categorization of behaviour: a comparison of dolphin whistle classification methods

VINCENT M. JANIK
School of Environmental and Evolutionary Biology, University of St Andrews, U.K.

(Received 10 February 1998; initial acceptance 2 April 1998; final acceptance 30 June 1998; MS. number: 5779)

The categorization of behaviour patterns into separate classes is crucial to the study of animal behaviour. Traditionally researchers have classified behaviour patterns through careful observation by eye. Recently this method has been increasingly replaced by computer methods. While the definition and fine scale analysis that can be achieved with computers is desirable, only a few studies have actually looked at how these methods perform in comparison with human observation. I compared the classification of bottlenose dolphin, *Tursiops truncatus*, whistles by human observers with the performance of three computer methods: (1) a method developed by McCowan (1995, *Ethology*, **100**, 177–193); (2) a comparison of cross-correlation coefficients using hierarchical cluster analysis; and (3) a comparison of average difference in frequency along two whistle contours also using hierarchical cluster analysis. The whistle sample consisted of 104 randomly chosen whistles from a group of four captive bottlenose dolphins recorded both during periods when one was separate from the rest of the group and while they all swam in the same pool. The sample contained five individual-specific signature whistles and several nonsignature whistles. Five human observers, without knowledge of the recording context, were more likely than the computer methods to identify signature whistles that were used only while an animal was isolated from the rest of the group. I discuss the limitations of methods commonly used for pattern recognition in communication studies. The discrepancies between methods show how crucial it is to obtain an external validation of the behaviour classes used in studies of animal behaviour.

A crucial step in any study of animal behaviour is division of the observed behaviour into separate categories. If those chosen have any relevance to the animal itself, a selective usage of these patterns according to some external variable should be observable. Examples of such a variable are a particular context or individual. Thus, if a category is used only in one particular context or by only one individual, it confirms the biological significance of this category. This is one of the most basic principles in animal behaviour research.

All classification methods include decisions by the investigator as to what parameters should be considered and how they should be weighted. The most common approach is the classification by human observers using their pattern recognition abilities. There are two main problems with this method. One is the issue of observer bias. If a researcher wants to confirm a chosen category by an external variable as described above it is important to ensure that the initial categorization was carried out without any knowledge of when or by whom a behaviour pattern was produced. Martin & Bateson (1986) and Milinski (1997) have provided excellent reviews of this problem and how to avoid it. The other problem is the reproducibility of a categorization method. Two human observers might weigh parameters differently in their pattern recognition and so come up with different categories. This problem can be avoided by using several observers to obtain a measure of observer agreement. If agreement is high, one can assume that the method is reproducible by others.

One disadvantage of classifications by human observers is that threshold values for categorizing the behaviour patterns are not clearly defined. Furthermore, small parameter differences that might be relevant to the animal could be missed by the human. With recent developments in computer technology, an increasing number of studies have started to use computers to obtain threshold values and look at possible subclasses of behaviour that are characterized by small parameter differences or investigate parameter variations within call types (e.g. May et al. 1988; Janik et al. 1994; Slabbekoorn & ten Cate 1997). This approach is very powerful particularly if behaviour patterns can be separated by looking at one or

Correspondence and present address: V. M. Janik, MS 34, Woods Hole Oceanographic Institution, Department of Biology, Woods Hole, MA 02543, U.S.A. (email: vjanik@whoi.edu).

more crucial parameters that are sufficient to describe the different behaviour types (e.g. in categorical perception of mouse pup ultrasounds, Ehret & Haack 1981; Ehret 1992).

But is visual observation always a less adequate method? Another way of using computers for pattern recognition is to develop a similarity measure. Examples of such measures are cross-correlation coefficients or differences in average values such as the mean sound frequency of a call. Still another approach is the application of computer-based neural network systems. However, these methods often do not perform as well in pattern recognition as humans do (see Khanna et al. 1997; Lippmann 1997). Furthermore, in the neural network approach the threshold values used to define a particular category are often difficult to retrieve from the program (e.g. Lehky & Sejnowski 1988). Thus, a researcher has to think carefully about which method to use in a study. This is particularly important with complex patterns. To date, only a few studies have compared different classification methods (Nowicki & Nelson 1990; Terhune et al. 1993; Lippmann 1997). However, such studies are important to assess how useful a particular method is and to aid in choosing the most appropriate one.

In this paper, I investigate the advantages and disadvantages of four methods for the classification of bottlenose dolphin, *Tursiops truncatus*, whistles. To assess how useful different methods are, a baseline is needed that defines which behaviour types are the right ones, that is, which types correspond closest to natural categories formed by the animal. One way of obtaining such an external validation of behaviour types defined by a researcher is by looking at their usage by the animal. If a behaviour type turns out to be used very selectively in only one context, it must closely resemble a natural behaviour category of the animal. In bottlenose dolphins, such a selective usage has been found for at least one whistle type in the repertoire of each individual. Janik & Slater (1998) used visual classification to define whistle types before looking at when they were used by the animals in their study. Their results showed that each of four very stereotyped whistle types defined by them was used almost exclusively by only one individual and only if it was isolated from other members of its group. Thus, they were able to show that visual inspection of frequency spectrograms is a valid method for recognizing at least one natural category in a dolphin's whistle repertoire. Janik & Slater's (1998) findings correspond closely to those of Caldwell et al. (1990) who termed the most common whistle type that is produced by an isolated individual its signature whistle. In this study I take a subset of the whistles recorded from the dolphins in Janik & Slater's study and compare the results of three computer-based methods with those obtained through visual classification by human observers. The computer-based methods are: (1) a method developed by McCowan (1995) that normalizes whistles in duration and uses principal component analysis and *k*-means cluster analysis; (2) a comparison of cross-correlation coefficients using hierarchical cluster analysis; and (3) a comparison of average differences in absolute frequency that also uses hierarchical cluster analysis. My aim was to compare how well these computer methods could identify the signature whistle types already known to be used almost exclusively when an individual was isolated, and thus validated as natural categories of behaviour.

METHODS

The Whistle Sample

The sample of dolphin whistles used for this study was a subset of 104 randomly chosen whistles from a total of 1323 whistles recorded from four bottlenose dolphins in January 1996 at the Zoo Duisburg, Germany. Figures 1 and 2 show the entire sample. The dolphin group consisted of an adult male, an adult female, a subadult male and a juvenile female. Recordings were made either while all the animals swam together in the same pool or while one animal had moved into a separate pool. These separations were not induced but occurred spontaneously in the daily behaviour of the animals. Each pool was fitted with one hydrophone (Dowty SSQ 904). Both hydrophones were recorded with the same recording level on separate tracks of a Marantz CP 430 tape recorder. To identify whether the solitary individual in the small pool or the rest of the group in the main pool produced a sound, I compared the sound intensity on both tracks of the tape recorder using SIGNAL (Version 3.0) sound analysis software (Engineering Design, Belmont, Massachusetts, U.S.A.). In such a set-up, the track that has the higher intensity indicates from where a sound comes. More detailed information on the holding facilities at the Duisburg Zoo and on recording conditions can be found in Janik & Slater (1998). To classify whistles into types, spectrograms were calculated (sampling rate: 50 kHz, Fast Fourier Transform size: 1024; time resolution: 20.5 ms; frequency resolution: 48.8 Hz; weighting function: Hanning window) and a line spectrogram of the fundamental frequency was extracted with the SIGNAL software as described in Janik et al. (1994). This method provides a line that represents the contour of the fundamental frequency of the whistle. Bottlenose dolphins often produce multiloop whistles in which separate whistles follow each other closely and occur together most of the time. For the analysis here each separate whistle from such multiloop whistles was considered on its own. Each whistle was given an identification number. These numbers are used purely to refer to a particular whistle in the sample and they have no further meaning.

Human Observer Classification

All 104 line spectrograms were printed on separate sheets and five observers were asked to classify calls independently by their shape. The observer who classified whistles in Janik & Slater (1998) was not used in this study. Spectrograms were presented in random order. All observers had extensive experience in classifying bird sounds but no experience with dolphin sounds. No information on recording context or caller identity was given

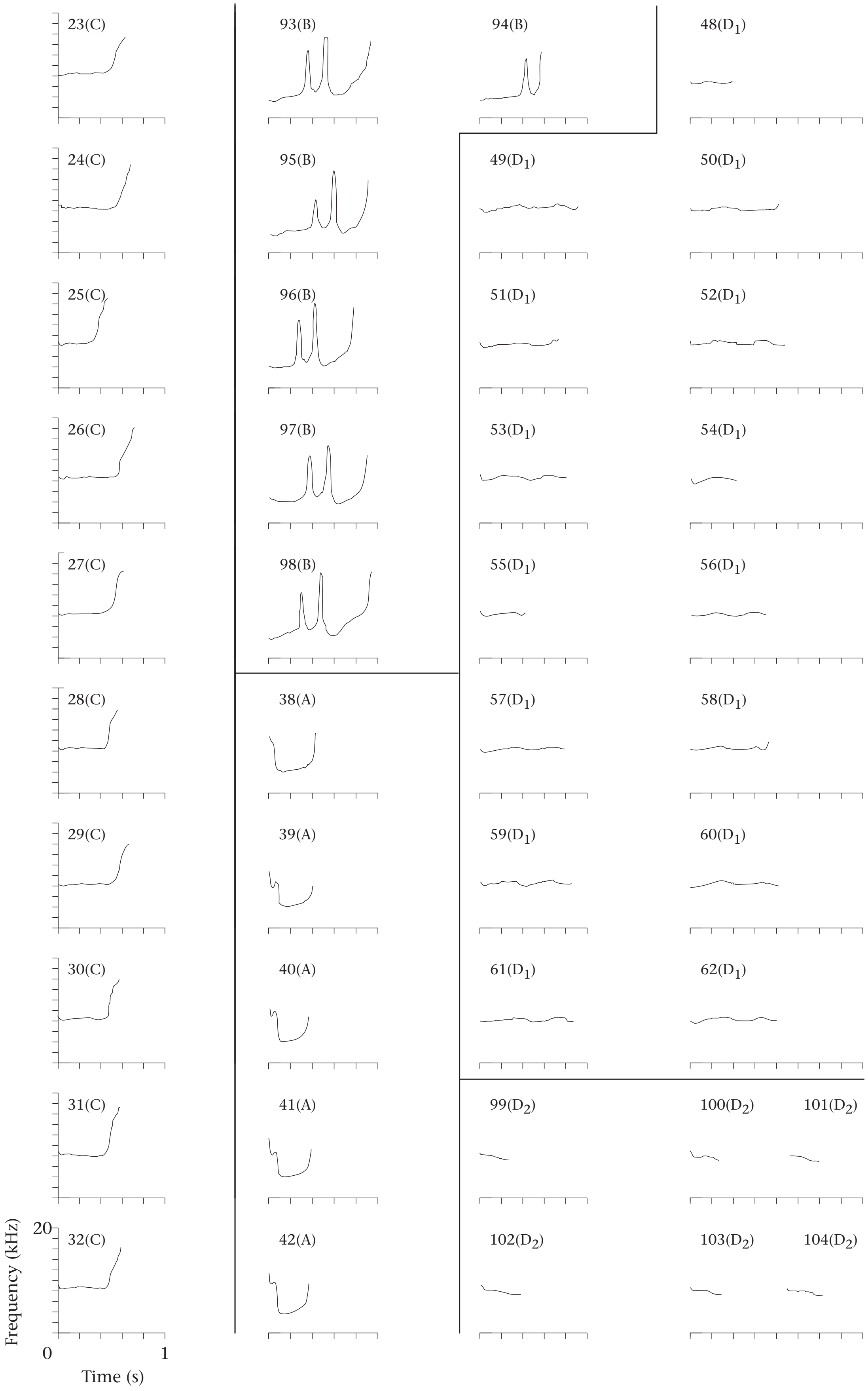


Figure 1. Line spectrograms of all signature whistles that were considered in this study (from Janik & Slater, 1998). The number on each spectrogram is its identification number followed by a letter indicating to which whistle type it belongs.

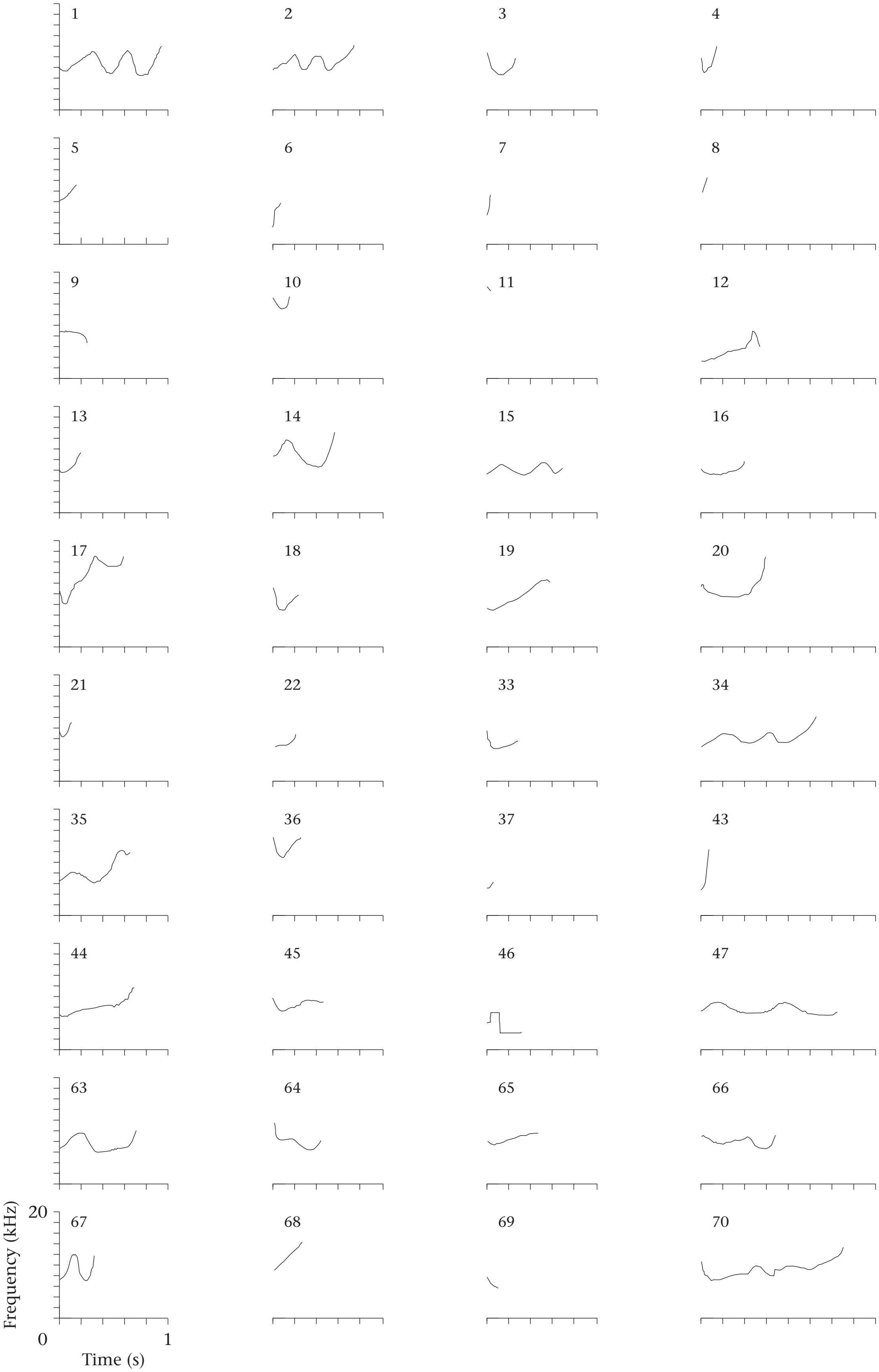

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
33
34
35
36
37
43
44
45
46
47
63
64
65
66
67
68
69
70
Frequency (kHz)
20
0
1
Time (s)

Figure 2. 1–70.

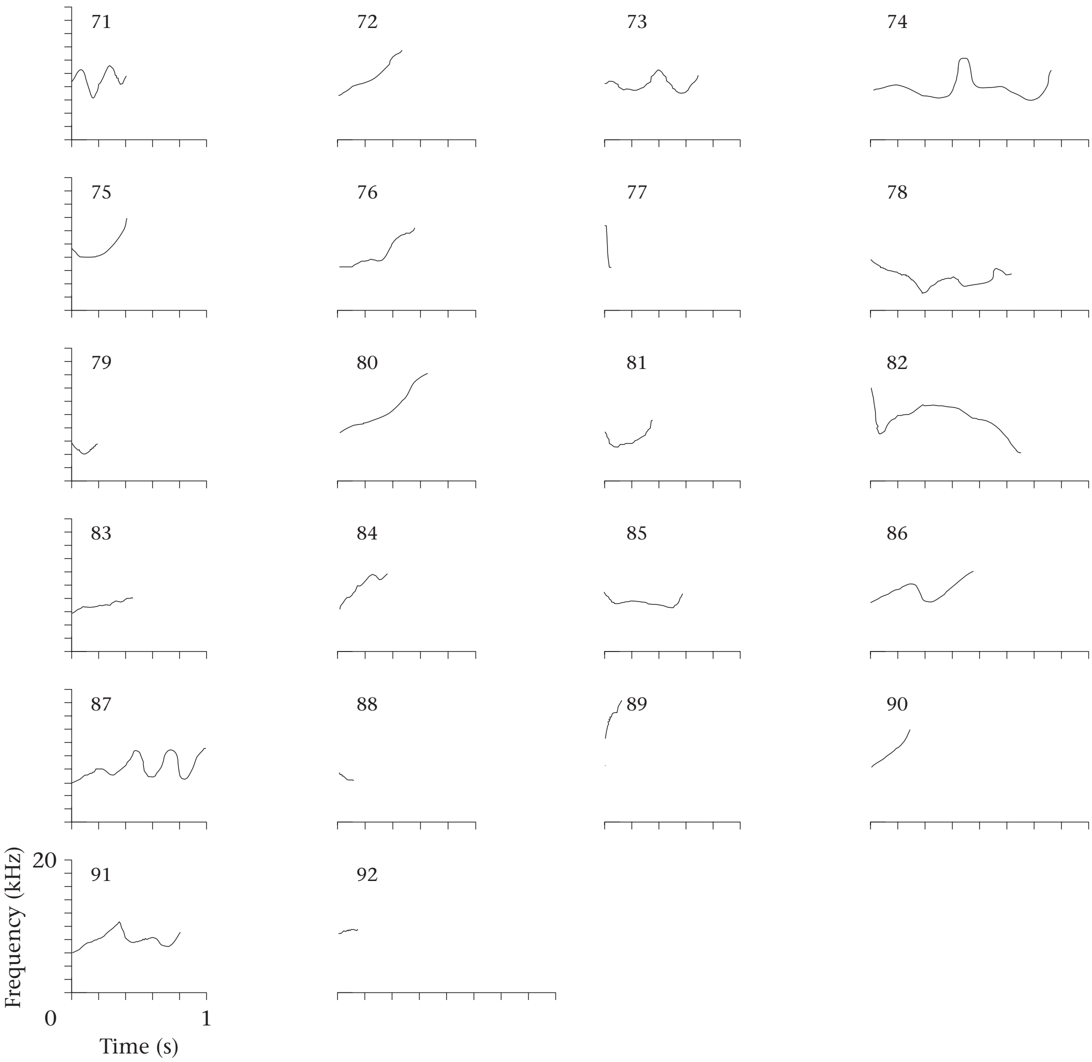


Figure 2. 71–92.

Figure 2. Line spectrograms of all nonsignature whistles that were considered in this study (from Janik & Slater 1998). The number on each spectrogram is its identification number.

to them. However, they were asked to pay particular attention to the possible occurrence of very stereotyped signals, so that they could be recognized and described as one type. Each observer was allowed to categorize the contours into as many classes as he or she thought appropriate. These types were then scanned for common types that could be found in the classification of all observers.

The McCowan Method

McCowan (1995) presented her own method to classify whistles. To replicate her method, 20 frequency measurements had to be taken from each whistle contour. These measurements were equally distributed over the contour by dividing the duration of each whistle by 19 and then taking frequency measurements at every 1/19th point of the total duration including the start and the end of the whistle. The resulting 20 frequency measurements were then taken as 20 variables for further calculations. This method eliminates any differences in the duration of whistles. All whistles are represented by the same number of frequency measurements and are, therefore, compared only by the shape of their contour.

In the next step the 20 variables were taken to compute a Pearson product–moment correlation matrix resulting in a similarity measure for each pair of whistles in the sample. A principal component analysis on the correlation matrix was carried out to reduce the number of collinear variables. Only factors with an eigenvalue of greater than 1.0 were used for subsequent analysis. In the final step factor scores from each data set of whistles were used in *k*-means cluster analyses using BMDP (Version 1988) statistical software, the package used by McCowan (1995). In a *k*-means cluster analysis the user specifies the number (*k*) of clusters to which cases should be allocated. After a random initial partition of the data set the centroid of each cluster is calculated. Then, each data point is allocated to the cluster with the centroid that is nearest to itself. Once all data points have been ascribed to a cluster, the centroids of these new clusters are calculated and each data point is reallocated again if it is closer to one of the new centroids of another cluster. These steps are repeated until no data point changes clusters anymore or a set number of iterations is reached. As in other cluster analysis techniques, the researcher has to decide which number of clusters corresponds to the actual structure in the data. This can be done by inspecting the results of several *k*-means cluster analyses in each of which the number (*k*) of clusters was different. McCowan (1995) used the cluster solution that produced the maximum number of nonoverlapping clusters as indicated by BMDP. However, BMDP only indicates overlap in a two-dimensional representation of a *k*-dimensional space (Dixon et al. 1990). Thus, clusters can overlap without BMDP indicating an overlap, or they can overlap in the two dimensions but be clearly separate in a dimension not displayed. The overlap indication was therefore not considered a satisfactory criterion to decide which cluster solution was appropriate. Instead I inspected all cluster solutions for $5<k<51$ for possible agreement in whistle classification with the other methods. All analyses were conducted using BMDP default settings (maximum iterations: 30; Dixon et al. 1990).

Cross-correlations and Cluster Analyses

Finally, I compared two different similarity measures and two different cluster analysis methods for their usefulness in whistle classification. I calculated the first similarity measure by cross-correlating every contour with all other contours in the sample. In this method the two contours were aligned so that the cross-correlation coefficient yielded its maximum. This maximum coefficient was then used as a similarity measure. Beeman (1996) and Khanna et al. (1997) give the formula for the calculation of the cross-correlation coefficient in this procedure. I used SIGNAL software for all cross-correlation analyses. The XCS command in SIGNAL was used to perform cross-correlations with a sliding time normalization. At each step in this procedure the shorter whistle is always just correlated with that part of the longer whistle with which it currently overlaps. The shorter of two whistles had to have at least 75% of the duration of the longer whistle. This threshold was set arbitrarily. Otherwise the cross-correlation coefficient was set to nil, the value for two very different contours.

The second similarity technique also involved cross-correlating contours. The two contours were aligned so that the cross-correlation yielded its maximum value as described above, but instead of using the correlation coefficient I calculated the absolute difference in frequency between the two contours every 5 ms. All differences were added up and then divided by the number of differences calculated. If one whistle was longer than the other the values were added only over the duration of the shorter whistle. Again the shorter whistle had to have at least 75% of the duration of the longer one. If the difference in duration was larger, the similarity value was set to 20 000, the value for two very different contours in this comparison.

Whistle type	A	B	C	D_1	D_2
Whistle ID numbers	**38** (3)	**93**	**23**	**48** (4)	69 (2)
	39	**94**	**24**	**49**	88 (2)
	40	**95**	**25**	**50** (4)	**99**
	41	**96**	**26**	**51** (4)	**100** (4)
	42	**97**	**27**	**52**	**101**
	81 (1)	**98**	**28**	**53**	**102**
			29	**54** (4)	**103**
			30	**55** (4)	**104**
			31	**56**	
			32	**57**	
			20 (1)	**58** (4)	
				59	
				60	
				61	
				62	
				65 (1)	
				83 (1)	

Figure 3. Human classification of dolphin whistles. The numbers correspond to the identification numbers of whistles in Figs 1 and 2. Numbers in parentheses indicate how many observers put the corresponding whistle into a type. No parentheses indicate that all five observers agreed on the classification of a whistle. Boxes indicate which whistles were signature whistles, that is, used by only one dolphin when separated from its group. Identification numbers of signature whistles are printed in bold.

These two methods resulted in two matrices, one with a measure of similarity (the cross-correlation coefficients) and the other with a measure of dissimilarity (the average frequency difference between all pairs of whistles). Each matrix was used for hierarchical cluster analyses for which I used the SPSS (Version 6.1) statistical software package using the between-groups average linkage method and the complete linkage method. The average linkage method is one of the most commonly used clustering methods in the biological sciences. It requires that a whistle has to be within a certain level of similarity to the average of the cluster to be included in that cluster. I compared it with the complete linkage method, which requires that a whistle has to be within a certain level of similarity to all members of that cluster. This latter method should favour the formation of very stereotyped whistle types (Aldenderfer & Blashfield 1984). I compared the results for all the methods used.

RESULTS

Signature Whistle Classification

The visual inspection method revealed that observers agreed on the classification of signature whistles. Five very stereotyped whistle types could be found in the classification of all observers (types A, B, C, D_1 and D_2; Fig. 3). Each of these types was used exclusively by only one individual dolphin if it swam isolated from its group members. In the larger sample of the Janik & Slater (1998) study only five of 439 such signature whistles were copies produced by other individuals and only 17 were produced while all animals swam in one pool. Following the definition by Caldwell et al. (1990) that the most

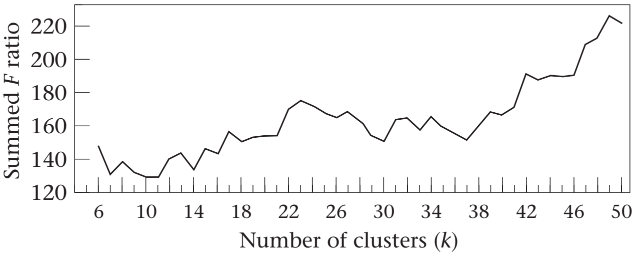
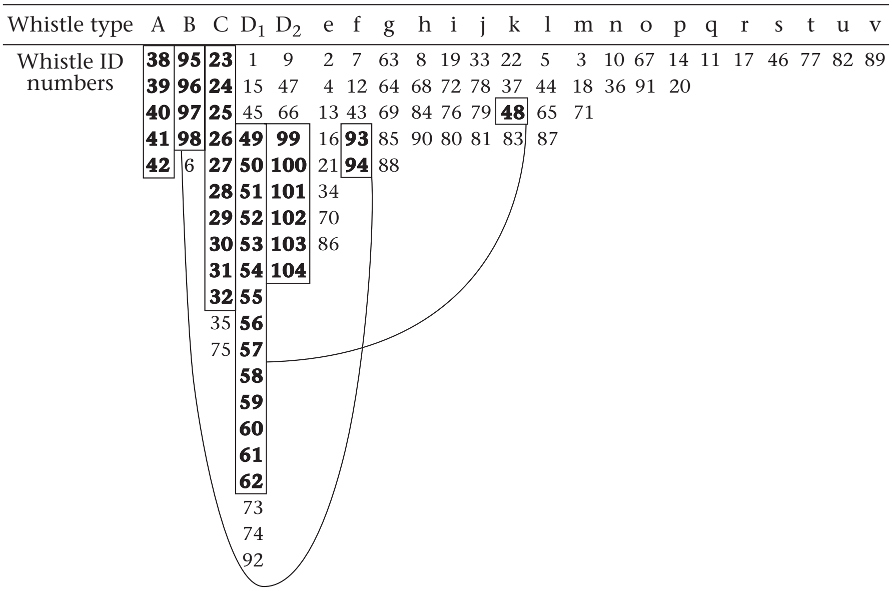


Figure 4. Distribution of summed *F* ratios from the McCowan method. A local peak can be found at *k*=23. The cluster solution of that point is shown in Fig. 5.

common whistle type of an isolated individual is its signature whistle, these five whistle types represented the signature whistles in the sample. Whistle types D_1 and D_2 were two parts of a multiloop signature whistle used by one of the animals. They occurred together most of the time in the study by Janik & Slater (1998). If only signature whistle types were considered, and all others were considered as a single residual class, observer agreement was extremely high (Kappa statistic (Siegel & Castellan 1988): κ=0.92, Z=22.37, P<0.0001). The observer classification was identical with the one in Janik & Slater (1998). However, Fig. 3 shows that up to two observers in this study sometimes also included one or two other whistles in a signature whistle type. Type D_1 was split into two types by one observer, but no additional nonsignature whistles were included with the resulting types. The observer indicated in his classification that he saw these two types as subtypes of one type. To be conservative only the type that had more whistles in it was considered in Fig. 3. Thus, there are a few whistles that only four observers agreed on in type D_1.

The principal component analysis on the 20 frequency measures taken from each whistle to reproduce McCowan's method resulted in three principal components with eigenvalues greater than 1.0. The *k*-means cluster analyses on the factor loadings of these components revealed that this method could not identify signature whistles as reliably as the human observers. Only one signature whistle type (type A) was reliably recognized in solutions where 9<*k*<43, where *k* is the number of clusters in the *k*-means cluster analysis. All whistles of type C were grouped together in all solutions where 6<*k*<26. However, in each solution between 2–13 other whistles were included in the same cluster as the type C whistles. A similar situation was found for type D_2 (found in solutions where 12<*k*<42, number of other whistles in the cluster from three to eight). Type D_1 was found in all solutions where 12<*k*<39, but it always had 6–11 other whistles in the same cluster and one of the D_1 signature whistles (number 48) was always placed in a separate type with several nonsignature whistles. Type B whistles were never all together in one cluster. The additional whistles found in signature whistle clusters were never classified as belonging to that cluster by human observers. They were also not produced by the respective individual in isolation in Janik & Slater's (1998) study. One method to select the best solution in *k*-means cluster analysis is a comparison of the sums of *F* ratios (between cluster sum-of-squares/within cluster sum-of-squares). The solution that maximizes the sum of the *F* ratios is then selected (Nowicki & Nelson 1990). Figure 4 shows the distribution of summed *F* ratios for the McCowan method. As the cluster number increases towards 50 clusters the summed *F* ratio increases. However, at *k*=23 a local maximum is reached. Figure 5 shows the classification at that point. Only two of the signature whistles (types A and C) that were found by the human observers were identified equally well by the McCowan method.



Whistle type	A	B	C	D_1	D_2	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v
Whistle ID numbers	**38**	**95**	**23**	1	9	2	7	63	8	19	33	22	5	3	10	67	14	11	17	46	77	82	89
	39	**96**	**24**	15	47	4	12	64	68	72	78	37	44	18	36	91	20						
	40	**97**	**25**	45	66	13	43	69	84	76	79	**48**	65	71									
	41	**98**	**26**	**49**	**99**	16	**93**	85	90	80	81	83	87										
	42	6	**27**	**50**	**100**	21	**94**	88															
			28	**51**	**101**	34																	
			29	**52**	**102**	70																	
			30	**53**	**103**	86																	
			31	**54**	**104**																		
			32	**55**																			
			35	**56**																			
			75	**57**																			
				58																			
				59																			
				60																			
				61																			
				62																			
				73																			
				74																			
				92																			

Figure 5. Whistle classification using the McCowan method (*k*=23). Each number represents one particular whistle (see Figs 1 and 2). Signature whistles are bold; those belonging to one type are boxed and, if split, connected by a curved line.

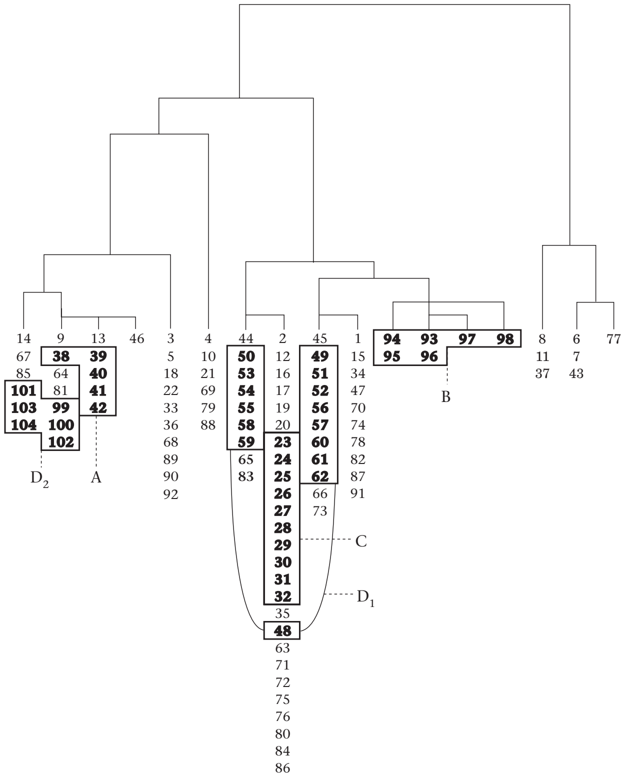


Figure 6. Cluster tree of the matrix of cross-correlation coefficients. Cluster method: between-group average linkage. Numbers are identification numbers of the whistles (see Figs 1 and 2). Signature whistles are bold; those belonging to one type are boxed and, if split, connected by arced lines. Capital letters indicate which box represents which signature whistle type. The cluster was drawn using rescaled distance measures.

Figure 6 shows the cluster tree that resulted from the analysis of the cross-correlation coefficients with average linkage cluster analysis. It becomes clear that this method can pick out one signature whistle type reliably only if cut at the right point (type B). Other signature whistle types were either grouped together with many other whistles or split into different clusters. Trees created with the two different cluster analysis methods were almost identical. The analysis of the average frequency differences between whistle contours was more successful in identifying signature whistles (Fig. 7). Here four out of five signature whistles could be identified. However, it depended again at what distance level in the tree would be used to define whistle types. The appropriate level was different for different signature whistle types. This frequency difference method was, like the McCowan method and the cross-correlation method, not suitable to identify signature whistles in the sample. Again, using the two different cluster analysis methods resulted in almost identical trees.

Classification of Nonsignature Whistles

Additional whistle types similar to those described in Tyack (1986) and Janik et al. (1994) could be found in the classifications of the five observers, but observer agreement was low. The following triplets of whistles were found together in every observer's classification: 8-43-89, 5-72-80, 18-33-36, 34-47-87 (Fig. 2). However, the number of different whistles with which they were grouped was large and varied between observers. In each case the whistles observers agreed on formed less than 60% of that type in each observer's category. More than three whistles were never put together by all observers.

Of the nonsignature whistles that showed the highest observer agreement in the human observer classification only two (numbers 72 and 80) were grouped together by the McCowan method. The method using cross-correlation coefficients, however, disagreed on only one of the triplets defined by the human observers (8-43-89) and did not group whistle number 5 with 72 and 80. Finally, the frequency difference method grouped some pairs of these whistles together, but found none of the triplets.

A comparison of the classification of nonsignature whistles between the computer methods revealed that they also showed very little agreement. But while the McCowan method resulted in very different whistle clusters, some of the differences between the cross-correlation coefficients and the frequency difference tree seemed to result from the finer resolution of the latter tree. None of the nonsignature whistle types defined by any of the methods was used selectively by only one individual or only in isolation.

DISCUSSION

The results showed clearly that methods agreed to only a very limited extent. Signature whistles could be identified by human observers but none of the computer methods was capable of identifying them reliably. It is important to note that only after the whistle types had been defined by humans was it found that these whistle types were used almost exclusively by one animal and only if it was isolated from its group (Janik & Slater, 1998). Even though it is unlikely that the perception of a whistle by a human observer maps exactly on to that of a dolphin, such an exclusive use of a behaviour type is rare. It shows that the human classification has recognized a class of behaviour that is significant for the animal. Such an external validation justifies the usage of a particular method if data on how the animal perceives and classifies whistles are not available. Such a justification is needed no matter whether the classification method is based on human observers or on a computer.

It is still possible that the computer methods, while failing to identify signature whistles reliably, could have discovered significant classes that were missed by the humans. It could be that dolphins use very different criteria for the classification of signature and nonsignature whistles. However, with the exception of the signature whistles, none of the classes defined by the computer methods was used exclusively by one or more animals in isolation. Thus, this context does not provide an external validation for any of those whistle types. Nevertheless, these types could be important in other contexts that

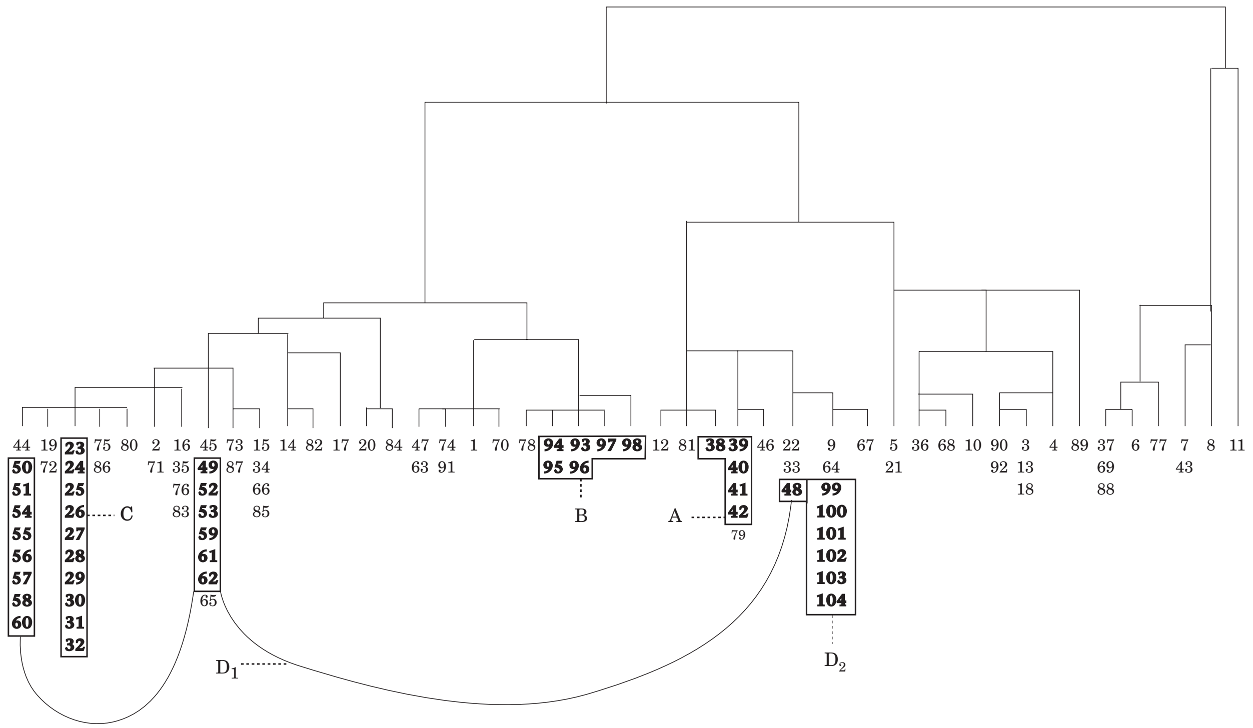


Figure 7. Cluster tree of the matrix of average frequency differences. Cluster method: between-group average linkage. Numbers are identification numbers of the whistles (see Figs 1 and 2). Signature whistles are bold; those belonging to one type are boxed and connected by arced lines. Capital letters indicate which box represents which signature whistle type. The cluster was drawn using rescaled distance measures.

were not considered in this study. However, the McCowan method on the one hand and the hierarchical cluster analysis techniques on the other did not agree on the classification of nonsignature whistles either. Furthermore, the signature whistles they split into several clusters were different ones. This disagreement between the computer methods showed that they concentrated on very different features of whistles. Studies on bottlenose dolphins in which McCowan's method has been used (McCowan & Reiss 1995a, b, 1997) did not find clear differential usage of whistle categories as this study has. It is difficult to assess what role the whistle types described in McCowan & Reiss's studies play in the animals' repertoires as an external validation for them is lacking. As mentioned above such an external validation has so far been achieved only for signature whistles (Janik & Slater 1998), but these whistle types could not be identified reliably by the McCowan method in this study.

The reasons for the inferior performance of the computer methods can be found in the way they compare whistle contours. If human observers compare whistles of types C and D_1, for example, they classify them as very different because the extended but very brief upsweep at the end of type C is lacking in type D_1. Thus, humans use the overall shape of the contour to classify them. The cross-correlation method and the average frequency difference method had difficulties separating these types because C and D_1 are very similar apart from the upsweep. These methods assess similarity over the whole contour and weigh each part of it equally. To them types C and D_1 are very similar since for most of the duration these types even overlap in their absolute frequency. Similarly, the average frequency difference method placed whistle number 73 close to type D_1 since they lie in the same frequency band. However, to the human observer number 73 has a clear wave shape while type D_1 is relatively flat. But the cross-correlation method and the frequency difference method also disagreed on how type D_1 should be split. The reason for this could be that the similarity value in the cross-correlation method is not linearly dependent on the slope of the frequency modulations in the comparison (Khanna et al. 1997) while this is the case in the frequency difference method.

The main problems for the McCowan method seemed to be its normalization for differences in duration and the number of frequency measurements taken from each contour. For example, one of the type D_1 whistles (number 48) was placed in a new type together with whistle number 37 even though their durations were very different. The relatively small number of frequency measurements seemed to be a problem for the classification of whistle type B. This whistle type is relatively long and has some rapid frequency modulations that cover a large frequency band. It is possible that the 20 equally distributed frequency measurements on each contour have sometimes missed these modulations so that the type was split into two. If both modulations in type B were missed the whistle would look like a simple upsweep. This, together with the fact that the McCowan method ignores differences in duration, could be the reason why whistle numbers 93 and 94 and numbers 7, 12 and 43 were placed into the same type.

The superiority of the human observer method in this study suggests that dolphins use the overall gestalt of a signature whistle to classify it. However, we really need perceptual studies on dolphin whistles to recognize the boundaries of their natural categories. Experiments that test how an animal would categorize a whistle (e.g. by conditioning an individual to discriminate between two whistle types and then testing how it classifies abbreviated whistles or whistles that have features of both types) are particularly useful in the study of natural categories and have been used successfully in other species (e.g. Horning et al. 1993). Such experiments are needed for signature whistles but also for nonsignature whistles. Observers and computer methods disagreed strongly on the classification of nonsignature whistles in this study. Previous studies have used general design features of whistles for the classification of nonsignature whistles, such as generally rising frequency, sinusoidal modulation or falling frequency contour (Tyack 1986; Janik et al. 1994). Such types could also be found in the classification by the observers used in this study, but the boundaries of these types were diffuse and the observers disagreed on borderline cases. Perception experiments should start by concentrating on gestalt perception, but should also try to assess the stability of whistle recognition if parameters such as duration start to vary. The McCowan method, for example, assumes that duration is irrelevant to the classification of whistles. To date, there is no evidence that this is the case. Bottlenose dolphins vary the duration of given whistle types according to the context (Janik et al. 1994). A certain stability of natural categories towards parameter changes can be assumed, but it is likely that there is a point at which whistle type identification by the animal starts to break down. Perception experiments could help us to understand, for example, whether a very short whistle is an interrupted version of a longer type or simply a complete short version of yet another whistle type.

Similar problems exist in the frequency domain. Bottlenose dolphins also vary frequency parameters in relation to context (Janik et al. 1994). But again it is likely that there are limits within which parameters have to be found for a whistle to be ascribed to a certain type by the animal. Richards et al. (1984), for example, suggested that bottlenose dolphins are not sensitive to the frequency band a signal lies in but only to its general shape. This is based on their finding that the experimental animal imitated an artificial low-frequency model sound but transferred it up one octave. However, it is premature to assume that absolute frequency is unimportant in the classification of dolphin whistles. Ralston & Herman (1995) showed that dolphins are able to learn to classify frequency contours of the same shape that lie in different frequency bands as one type. However, their study animal concentrated on absolute parameter differences in its classification in the initial stages of the training when it put such whistles into different types.

The issues I have discussed here using the example of whistle classification in dolphins are relevant to all observations of animal behaviour. Computer methods are widely used to classify behaviour patterns. The comparison in this study showed that researchers have to be careful that their chosen method identifies patterns that are relevant to the animal. The use of a computer method is desirable for many reasons, but it has to be tailored carefully towards the biological question that is investigated.

Acknowledgments

I thank the staff of the Zoo Duisburg for the opportunity to work with their animals and for their support during this project, especially Roland Edler, Director Reinhard Frese, Manuel García Hartmann, Friedrich Ostenrath and Ulf Schönfeld. The help of Diego Gil, Alex Jones, Nigel Mann, Katharina Riebel and Peter Slater with the whistle classification was also greatly appreciated. Rob Marshall (Stirling University) and Fiona O'Callaghan (BMDP software) gave valuable advice and help concerning the BMDP software. Alex Jones, Katharina Riebel and Peter Slater provided valuable comments on the manuscript. I also thank Peter Tyack for discussions on whistle classification. This research was funded by a DAAD Doktorandenstipendium aus Mitteln des zweiten Hochschulsonderprogramms.

References

Aldenderfer, M. S. & Blashfield, R. K. 1984. *Cluster Analysis*. Newbury Park: Sage.

Beeman, K. 1996. *SIGNAL User's Guide, Version 3.0*. Belmont, Massachusetts: Engineering Design.

Caldwell, M. C., Caldwell, D. K. & Tyack, P. L. 1990. Review of the signature-whistle-hypothesis for the Atlantic bottlenose dolphin. In: *The Bottlenose Dolphin* (Ed. by S. Leatherwood & R. R. Reeves), pp. 199–234. San Diego: Academic Press.

Dixon, W. J., Brown, M. D., Engelman, L. & Jennrich, R. I. 1990. *BMDP Statistical Software Manual*. Berkeley, California: University of California Press.

Ehret, G. 1992. Categorical perception of mouse-pup ultrasounds in the temporal domain. *Animal Behaviour*, **43,** 409–416.

Ehret, G. & Haack, B. 1981. Categorical perception of mouse pup ultrasounds by lactating females. *Naturwissenschaften*, **68,** 208.

Horning, C. L., Beecher, M. D., Stoddard, P. K. & Campbell, S. E. 1993. Song perception in the song sparrow: importance of different parts of the song in song type classification. *Ethology*, **94,** 46–58.

Janik, V. M. & Slater, P. J. B. 1998. Context-specific use suggests that bottlenose dolphin signature whistles are cohesion calls. *Animal Behaviour*, **56,** 829–838.

Janik, V. M., Dehnhardt, G. & Todt, D. 1994. Signature whistle variations in a bottlenosed dolphin, *Tursiops truncatus*. *Behavioral Ecology and Sociobiology*, **35,** 243–248.

Khanna, H., Gaunt, S. L. L. & McCallum, D. A. 1997. Digital spectrographic cross-correlation: tests of sensitivity. *Bioacoustics*, **7,** 209–234.

Lehky, S. R. & Sejnowski, T. J. 1988. Network model of shape-from-shading: neural function arises from both receptive and projective fields. *Nature*, **333,** 452–454.

Lippmann, R. P. 1997. Speech recognition by machines and humans. *Speech Communication*, **22,** 1–15.

McCowan, B. 1995. A new quantitative technique for categorizing whistles using simulated signals and whistles from captive bottlenose dolphins (Delphinidae, *Tursiops truncatus*). *Ethology*, **100,** 177–193.

McCowan, B. & Reiss, D. 1995a. Quantitative comparison of whistle repertoires from captive adult bottlenose dolphins (Delphinidae, *Tursiops truncatus*): a re-evaluation of the signature whistle hypothesis. *Ethology*, **100**, 194–209.

McCowan, B. & Reiss, D. 1995b. Whistle contour development in captive-born infant bottlenose dolphins (*Tursiops truncatus*): role of learning. *Journal of Comparative Psychology*, **109,** 242–260.

McCowan, B. & Reiss, D. 1997. Vocal learning in captive bottlenose dolphins: a comparison with humans and nonhuman animals. In: *Social Influences on Vocal Development* (Ed. by C. T. Snowdon & M. Hausberger), pp. 178–207. Cambridge: Cambridge University Press.

Martin, P. & Bateson, P. 1986. *Measuring Behaviour*. Cambridge: Cambridge University Press.

May, B. J., Moody, D. B. & Stebbins, W. C. 1988. The significant features of Japanese monkey coo sounds: a psychophysical study. *Animal Behaviour*, **36,** 1432–1444.

Milinski, M. 1997. How to avoid seven deadly sins in the study of behavior. *Advances in the Study of Behavior*, **26,** 160–180.

Nowicki, S. & 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.

Ralston, J. V. & Herman, L. M. 1995. Perception and generalization of frequency contours by a bottlenose dolphin (*Tursiops truncatus*). *Journal of Comparative Psychology*, **109,** 268–277.

Richards, D. G., Wolz, J. P. & Herman, L. M. 1984. Vocal mimicry of computer-generated sounds and vocal labeling of objects by a bottlenosed dolphin, *Tursiops truncatus*. *Journal of Comparative Psychology*, **98,** 10–28.

Siegel, S. & Castellan, N. J., Jr 1988. *Nonparametric Statistics for the Behavioral Sciences*. 2nd edn. New York: McGraw-Hill.

Slabbekoorn, H. & ten Cate, C. 1997. Stronger territorial responses to frequency modulated coos in collared doves. *Animal Behaviour*, **54**, 955–965.

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

Tyack, P. 1986. Whistle repertoires of two bottlenosed dolphins, *Tursiops truncatus*: mimicry of signature whistles? *Behavioral Ecology and Sociobiology*, **18,** 251–257.