# TEMPORAL SEQUENCE ANALYSIS OF BOTTLENOSE DOLPHIN VOCALIZATIONS

BY
JOSHUA T. ABBOTT

A Thesis

Submitted to the Divisions of Natural Sciences and Social Sciences
New College of Florida
in partial fulfillment of the requirements for the degree
Bachelor of Arts
Under the co-sponsorship of Karsten Henckell and Heidi Harley

Sarasota, Florida
May, 2009

# ACKNOWLEDGEMENTS

It's been quite a long journey through different undergraduate institutions, time off in the work force, and combining two majors to finally finish my bachelors degree. I thank my family for suffering through it all alongside me, and all the while giving constant encouragement and support for my decisions.

I thank the many friends I've made along the way: Steve, Sophia, Max, Phyllis, Tyler, Pete, Anwar, Renata, Indra, Ray, Shelly, Chad, Leah, and many, many others.

I thank the women who shaped me along the way: Corrie, Julia, Jess, and Genevieve.

I also thank the lab assistants and researchers who helped greatly with this study: Wendi Fellner, Jenna Clark, and Caitlin O'Brien.

And finally, I thank my committee: Dr. Heidi Harley, Dr. Karsten Henckell, and Dr. Pat McDonald. In particular, Heidi stood by me through the sleepless nights as new methods revealed themselves a mere three weeks before the thesis was due. I cannot thank her enough for this support.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ..... ii

LIST OF TABLES AND FIGURES ..... v

ABSTRACT ..... vi

CHAPTER 1: INTRODUCTION ..... 1

Categorization ..... 1

A Quantitative Measure of Similarity ..... 3

The Contour Similarity Technique ..... 6

A Comparison of Different Whistle Categorization Methods ..... 7

ARTwarp ..... 10

Information Theory ..... 12

Zipf's Law ..... 13

Entropy ..... 16

Markov Chains ..... 18

Statistical Analysis of Indus Scripts ..... 20

Introduction ..... 20

Cumulative Frequency Distributions ..... 21

Bi-gram Probabilities ..... 22

Log-Likelihood Significance Test ..... 23

Entropy and Mutual Information ..... 24

n-gram Modeling ..... 25
Conclusion ..... 26
CHAPTER 2: RESEARCH METHODS ..... 28
Subjects / Facility ..... 28
Stimulus ..... 29
Procedures ..... 29
Analysis Procedures ..... 30
CHAPTER 3: RESULTS ..... 32
What level of categorization is suitable for finding patterns? ..... 32
What is the distribution of sequence lengths in the dataset? ..... 34
How are the categories distributed in the dataset? ..... 35
What is the strength of correlation between consecutive vocalizations? ..... 36
CHAPTER 4: GENERAL DISCUSSION ..... 39
Interpretation of Analysis ..... 39
Future Research ..... 40
Conclusion ..... 41
REFERENCES ..... 42
APPENDIX I: CATEGORIZATION PROTOCOL ..... 46

LIST OF TABLES AND FIGURES

Figure 1. Example spectrograms of a burst pulse and a whistle……. 2

Figure 2. A comparison of contours before and after time-warping……... 4

Figure 3. An example spectrogram of a chick-a-dee call sequence…….. 15

Figure 4. An example Indus text. The sign script is located on the top…… 21

Figure 5. An illustration of the facility housing the dolphins under study……... 28

Figure 6. Distribution of sequences by length. Six sequences are not represented in this figure because they were more than 11 units long……. 34

Figure 7. Cumulative frequency distributions of all categories, sequence beginners, and sequence enders……. 35

Figure 8. Transition probabilities from the dataset. The top matrix represents an absence of correlation. The bottom matrix represents the distribution of the dataset….. 38

Table 1. The entropy and mutual information of the first dataset…… 32

Table 2. The entropy and mutual information of the second dataset…… 33

Table 3. The 10 most frequent vocalization pairs and 10 most statistically significant... 36

TEMPORAL SEQUENCE ANALYSIS OF BOTTLENOSE DOLPHIN VOCALIZATIONS

Joshua T. Abbott

New College of Florida, 2009

## ABSTRACT

This investigation explores the usefulness of statistical language processing methods for analyzing structure in sequences of vocalizations produced by Atlantic bottlenose dolphins (*Tursiops truncatus*). The inherent difficulties underlying such a process are examined and a suite of methods is established for the desired analysis. To date, few studies have addressed all the categories of dolphin vocalizations within a single analysis or the sequences in which they are produced. A protocol was developed to identify categories spanning the dolphin vocal repertoire and to define sequences of vocalizations. Varying datasets were constructed from an identified corpus to discriminate between differences in potential information content, based on assigning vocalizations to a few broad categories versus more finely defined groups. Sequences of broadly categorized vocalizations appeared to be randomly structured, but inclusion of finer groupings revealed a syntax; statistical analysis of a finely grouped dataset indicated the presence of correlations between successive vocalizations in a sequence. A well defined set of vocalizations also began and ended the sequences. Finally, analysis

revealed that a number of vocalization types disregarded in previous studies occurred at very high frequency counts and in statistically significant pairs of vocalizations as computed by a log-likelihood measure of bi-gram association. Altogether these data suggest that sequences of dolphin vocalizations are constructed in a non-random fashion, i.e., they follow syntactic rules.

Karsten Henckell
Division of Natural Sciences

Heidi Harley
Division of Social Sciences

# Chapter 1: INTRODUCTION

The current body of knowledge concerning bottlenose dolphins (*Tursiops truncatus*) has shown they are rather remarkable creatures. They can mimic computer-generated sounds (Richards, 1986), hear sounds over from 2 to 120kHz (Au, Popper, & Fay, 2000), apparently learn from other dolphins to use sponges as foraging tools (Kruzen, Mann, Heithaus, Conner, Bejder, & Sherwin, 2005), and understand sequence-sensitive instructions conveyed by imperative sentences produced with the grammar of an artificial acoustic language (Herman, Wolz, & Richards, 1984). They also have the highest encephalization quotient of all modern mammals except humans (Marino, 1996),

Dolphin's vocal emissions are complex: varied and produced across a very wide frequency range. To date, this system, though well studied in specific areas, has largely eluded a comprehensive analysis. Here we explore the tools required to conduct such an investigation and begin to apply them to a corpus of dolphin vocalizations. This chapter will address issues related to the underlying problems of temporal sequence analysis on dolphin vocalizations. It covers previous work on categorizing dolphin vocalizations, an issue central to determining the units to be extracted before sequence analysis can occur. In addition, the use of information theory to assess the complexity of an animal's vocal repertoire and a framework of statistical methods to infer syntax from an ancient human system of signs are also considered.

*Categorization:*

Dolphins produce vocalizations that can be divided into narrowband (energy in a single frequency at one point in time) whistles and broadband (energy in many frequencies at one point in time) echolocation clicks and burst pulses. Echolocation clicks

occur in trains and appear to be primarily used for navigational and object identification purposes, although some studies suggest social utilization as well (e.g., Xitco & Roitblat, 1996). Burst pulses are primarily thought to have a social function, however, they have not been extensively studied (Overstrom, 1983). Whistles have been the most commonly studied social dolphin vocalization and are primarily used for communication (Harley, 2008; Herman & Tavolga, 1980). Figure 1 gives example spectrograms of each of the previous vocalizations. A spectrogram is the most commonly used method of visually representing vocalizations. Essentially, it is a three dimensional graph with a horizontal axis of time, a vertical axis of frequency (usually in kHz), and a third dimension denoted by color for conveying signal intensity.

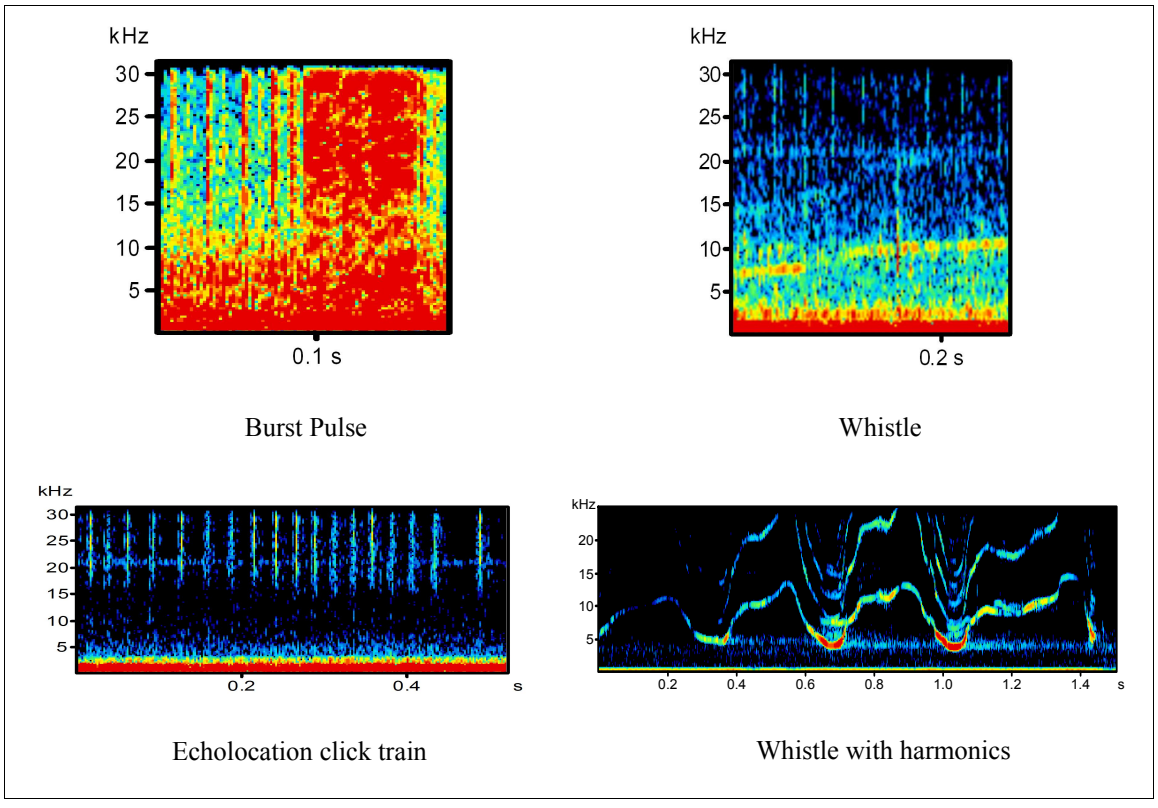



Figure 1. Example spectrograms of a burst pulse, click train, and whistles.

In 1965, Caldwell and Caldwell reported that dolphins appeared to produce distinctive whistle frequency contours, some of which are unique to each dolphin. The Caldwells proposed that these unique whistles were used by dolphins to transmit their identity and thus suggested the term "signature whistle" to describe these signals. The Caldwells also hypothesized that the contour of the fundamental frequency was a distinguishing characteristic for signature whistles. They observed that the overall shape of this contour generally remained consistent for each dolphin, although the whistle frequency, duration, and intensity varied across behavioral contexts. The Caldwells' discovery led to a plethora of work on signature whistles for the next several decades (Harley, 2008), including automated methods to find them.

*(i) A Quantitative Measure of Similarity*

One of the first studies to automate the process of comparing similarity in whistle contours, specifically for that of designating signature whistles, was produced by Buck and Tyack in 1993. After the Caldwells proposition of signature whistles, most work on distinguishing these whistles was based on qualitative measures of similarity between the contours of whistles. Human observers were quite good at this due to their pattern recognition abilities. However, there are some problems with this method, since using humans is slow and labor intensive. Buck and Tyack proposed that a quantitative measure of similarity between signature whistle contours would provide a faster, objective, and easily repeatable basis for categorizing whistles.

Dolphins produce whistles with fundamental frequencies usually in the human audible range (below 20 kHz). The fundamental frequency of a whistle is defined as the lowest frequency produced and is most often the strongest component of the signal.

Whistles often also have harmonics, which occur at integer multiples of the fundamental and often extend beyond the range of human hearing. Although signature whistle contours are stable, the whistles can vary in absolute parameters like amplitude and duration. When using an algorithm to compare whistle contours for similarity, there is an inherent problem of how to compensate for these variations in duration. Methods from speech recognition were borrowed to solve this problem and properly align the features of contours.

Buck and Tyack developed an algorithm to extract the fundamental frequency contour from a spectrogram and to perform dynamic time-warping on the associated contour extractions. The contour extraction component of the algorithm searched a spectrogram for the strongest components and recorded the frequency measurements for some given sampling frequency. The dynamic time-warping component of the algorithm was borrowed from Sakoe & Chiba (1978) with slight modification. Time-warping is a way to compute the minimum distance between two varying length vectors, usually representing a time series. As seen in Figure 2, time-warping can align contours varying in small features including duration.

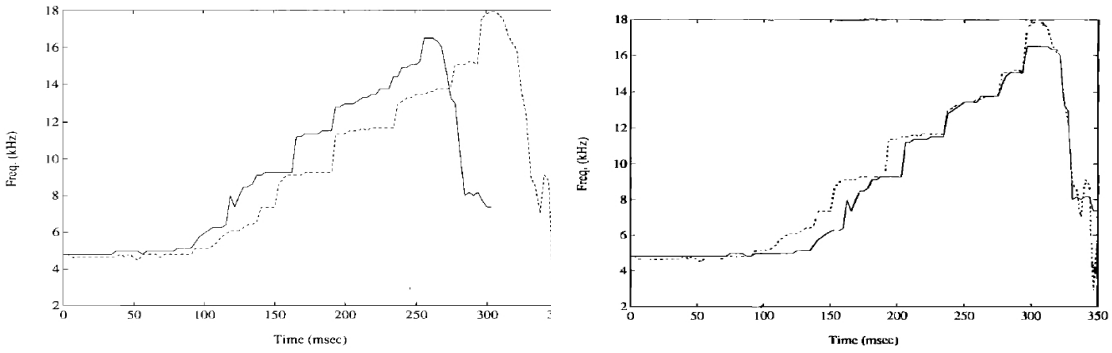



Figure 2. A comparison of contours before and after time-warping.

The graph on the left contains contours of two whistles produced by the same dolphin prior to time-warping. Note that the contours are not exactly identical, yet very similar in shape. The graph on the right shows the alignment of contours after warping. In this instance, the dashed line was held fixed while the solid line was time warped. While the alignment is not perfect, it is clear the features are visibly better aligned after warping.

To evaluate the efficacy of the similarity judgments, Buck and Tyack tested their algorithm's performance on out-of-set recognition of two sets of bottlenose dolphin whistles. Out-of-set recognition essentially assigns an unknown signal to some known reference signal out of a set of possible reference signals through a distance metric. Buck and Tyack's system was able to match correctly all unknown whistles with the dolphin that produced them. The main conclusion from this work was that the fundamental frequency contour is a primary key in individual identification for bottlenose dolphin signature whistles and that dynamic time-warping algorithms are well suited for categorization of signature whistles.

This study has influenced researchers classifying vocalizations of killer whales as well. Testing multiple dynamic time-warping algorithms, Judith Brown and others were able to create automated classification schemes that achieved very high agreement with human classifiers (Brown, Hodgins-Davis & Miller, 2006; Brown & Miller, 2007). They found that the best results were generated by the algorithm of Sakoe and Chiba (Sakoe & Chiba, 1978), also used by Buck and Tyack.

*(ii) The Contour Similarity Technique*

In 1995, McCowan created her own method to classify whistles, which she named "the Contour Similarity Technique" (McCowan, 1995; McCowan & Reiss, 1995). Twenty equally distributed measurements from a whistle contour were entered into a Pearson product-moment correlation matrix to get a similarity measure for each pair of whistles in the sample. Afterwards, principle component analysis was conducted on the correlation matrix to reduce the number of collinear variables. Finally, the resulting factor scores from each data set of whistles was used as input in a *k*-means cluster analysis.

*K*-means cluster analysis uses an algorithm to put *n* data points in an I-dimensional space into *k* clusters, of which the user supplies *k* as input. When the algorithm begins, there are *k* means, or centroids, that are initialized to random values. The algorithm then iterates over two steps: an assignment step in which each data point is assigned to the nearest mean, and an update step in which the means are adjusted to match the sample means of the data points for which they are responsible. These steps are repeated until the assignments do not change (MacKay, 2003). The mean to which each data point is eventually assigned is the cluster to which it belongs.

As in other cluster analysis techniques, the researcher has to decide which number of clusters corresponds to the actual structure in the data. This decision can be determined by inspecting the results of several *k*-means cluster analyses, varying *k*. McCowan used the solution that produced the maximum number of non-overlapping clusters as indicated by BMDP, the statistical software package used in her study. However, BMDP only indicates overlap in a two-dimensional representation of a *k*-dimensional space (Dixon, 1990). Thus clusters can overlap without BMDP indicating an overlap, or they can

overlap in the two dimensions but clearly be separate in a dimension not displayed. The overlap indication was therefore not a satisfactory criterion to decide which cluster solution was appropriate (Janik, 1999).

Other problems with the Contour Similarity Technique stem from the use of 20 equally distributed frequency measurements and *k*-means cluster analysis itself. Since the initialization of the means is random, different runs with the same number *k* as input can result in different clusters. Thus there is not a sufficient condition to choose among the multiple alternative clusterings for a given *k*. Also, McCowan's method was utilized with another data set (Janik, 1999), and the small number of frequency measurements had poor categorization results on whistles that were relatively long and had some rapid frequency modulations. Finally, the method assumes that duration is irrelevant to the classification of whistles, yet, 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, Dehnhardt, & Todt, 1994).

McCowan and Reiss tried to fix some of these issues in a later study (McCowan & Reiss, 2001) by using 60 equally distributed frequency measurements of a whistle contour rather than the previous 20, and by expanding their whistle corpus. However, the flaws of simple *k*-means cluster analysis and the assumption that duration is irrelevant to the classification of whistles remained.

*(iii) A Comparison of Different Whistle Categorization Methods*

Categorization of whistles both by humans through visual inspection of spectrograms and by computational methods produces variable results. In 1999, Janik compared the categorization of dolphin whistles by human observers with the

performance of three computer methods: (1) McCowan's method, (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 results showed that the different methods of categorization agreed only to a very limited extent. Signature whistles could be identified by human observers but none of the computer methods was capable of identifying them reliably.

For the human observer categorization, 104 randomly chosen line spectrograms produced by four captive dolphins were printed on separate sheets and five observers were asked to classify calls independently by their shape. Any whistle separated by a pause, including multi-loop whistles, was considered an individual whistle. All observers had extensive experience in classifying bird sounds but no experience with dolphin sounds. No information on recording context or number of dolphins was given, but observers 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. Observers were also told to categorize the contours into as many classes as he or she deemed appropriate.

The McCowan method of categorization was the same as in McCowan and Reiss (1995), except that Janik inspected all cluster solutions in a finite range of *k* values for possible agreement in whistle classification to fix the previously discussed overlapping problem. The other two computer methods were cross-correlation techniques sensitive to duration differences. One method cross-correlated every contour with all other contours in the sample. The two contours being compared were aligned so that the cross-correlation coefficient yielded its maximum, which was then used as a similarity

measure. The second method was identical to the first, except instead of using the maximum coefficient, Janik calculated the absolute difference in frequency between the two contours every 5ms. All differences were added up and then divided by the number of differences calculated. These two methods resulted in two matrices, one with a measure of similarity via cross-correlation coefficients, and the other with a measure of dissimilarity via average frequency differences taken every 5ms between all pairs of whistle contours. Each matrix was then used for hierarchical cluster analyses using SPSS statistical software package and the between-groups average linkage and complete linkage methods. The average linkage method is one of the most commonly used clustering methods in biological sciences.

In comparing the results for all methods used, Janik found many discrepancies between the different categorizations. Human categorizers were best at categorizing whistles by whistler and context; agreement was very high for the predominant whistles in the data set. Humans appeared to use the overall shape of the contour, while the computer methods assessed similarity over the whole contour and weighed each part of it equally. The computer methods were not able to identify signature whistles reliably, and it is unknown whether the categories they did discover were relevant to dolphins. That humans sorted by whistler is an external validation of this sorting method because it acknowledges a category that is clearly significant for the animal. Such a justification is needed no matter whether the classification method is based on human observers or on a computer.

*(iv) ARTwarp*

The comparison of categorization methods in Janik (1999) showed that researchers have to be careful that their chosen method identifies patterns that are relevant to the animal under study. 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. Janik had this in mind when he and Deecke created a new method of categorization, called ARTwarp, that incorporated dynamic time warping and an adaptive resonance theory neural network (Deecke & Janik, 2006). The two researchers argue that previous automated classification schemes perform poorly due to failure in considering two fundamentals of acoustic perception when measuring the similarity of sound patterns: flexibility in the time domain and the exponential perception of sound frequency.

Flexibility in the time domain had been addressed previously by Buck and Tyack (Buck & Tyack, 1993), and ARTwarp utilizes their dynamic time-warping algorithm (with minor modifications) to allow for variation in the lengths of different components of the whistle contours. The other main perceptual feature to consider is that tonal frequency is not perceived on a linear scale but on a logarithmic scale. Humans perceive two tones with frequencies that differ by an octave as being the same. This perception is reflected in part by the distribution of hair cells sensitive to different frequencies in the inner ear. Thus, 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. ARTwarp accounts for this logarithmic scale by expressing similarity of contours as their relative similarity in frequency.

To categorize the whistle contours after warping, ARTwarp uses an adaptive resonance theory neural network based on the ART2 learning algorithm(Carpenter & Grossberg, 1987). ART2 is an unsupervised learning algorithm in which a given input pattern is compared to a subset of reference patterns for the categories found in the current run. If the input pattern resembles one of the reference patterns within a defined degree of similarity, deemed the "vigilance", the input is assigned to the category represented by this reference pattern, and the associated set of reference patterns is updated to account for 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 networks have the advantage that they do not require assumptions about the frequencies of patterns in different categories. Thus they lend themselves quite well to the categorization of whistle contours, where equal distribution cannot be assumed.

The set of dolphin whistle contours to be categorized was the same set of 104 calls used in Janik (1999). The vigilance level was obtained by categorizing the signature whistles of a sole individual and increasing the vigilance in steps of 1% until the analysis split these signature whistles into different categories. A critical vigilance of 96% was the highest value that still maintained the whistles in 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.

On this corpus of 104 whistle contours, ARTwarp was able to recognize biologically meaningful categories even though it was not designed to detect individual

signature whistles and identify them as such. It recognized the stereotyped signature whistles to a high degree and thus performed much better than any of the computational methods tested in Janik (1999). ARTwarp even performed marginally better at detecting the signature whistle categories in the given data set than the human observers in the previous study (Janik, 1999). However, this method did not agree with the human observers in the categorization of non-signature whistles. Since there is no external validation for appropriate classification of non-signature whistles, it is impossible to say which categorization scheme is of greater biological relevance for these other whistles.

*Information Theory*:

Aside from research into placing whistles into relevant categories, there has also been work on modeling the complexity of the whistle repertoire. To formalize complexity, researchers have attempted to utilize the field of information theory to examine the channel capacity and structure of animal communication systems (Hailman, Ficken, & Ficken, 1985; McCowan, Hanser, & Doyle, 1999, 2002; Suzuki, Buck & Tyack, 2006). Information theory originated in the study of electrical communication systems but provides an abstract framework for measuring different types of information transmission (Shannon, 1948). The technical concept of information denotes how much choice one has when producing a signal sequence from some set of possible symbols. The capacity of a signal generating process to encode information is measured by Shannon's equation for entropy:

$$E = \sum_{i}^{n} p_i \log_2(1/p_i)$$

where $p_i$ is the probability of occurrence of the $i^{th}$ symbol, $n$ is the number of possible symbols, and $E$ is measured in bits/symbol. This section reviews three methods of analysis derived from the concept of information as a metric of signal complexity.

*(i) Zipf's Law*

In McCowan, Hanser, & Doyle's analysis of dolphin whistles (1999), the researchers argue that although information theory cannot directly measure the type of information transmitted, it can indirectly give insight into signal meaning because the amount of information is often linked to the type of information being transmitted. In particular, McCowan et al. utilized Zipf's law as a measure to assess the structural composition of a dolphin's whistle repertoire. Zipf's law is an empirical rule stating the number of occurrences of a word in a large corpus of text is inversely proportional to the order of frequency of occurrence (Pierce, 1980). For example, the hundredth most frequent word should occur approximately 1/100 as many times as the most frequent word.

Using the same data set from 1995, categorized by the Contour Similarity Technique, McCowan and her colleagues regressed the logarithm of the rank of a dolphin whistle (the ordering of the number of times a whistle occurred) on the logarithm of their actual frequency of occurrence. In Zipf's original paper, he found that many different human languages had a regression slope of approximately -1.00. This balance was proposed to optimize the communicative capacity because the structure of the system is neither too repetitive nor too diverse (Zipf, 1949). McCowan found her data set had a regression slope of -0.95, which she believes indicates "that the distribution of whistles in

the dolphin's repertoire is indeed non-random and, in fact, closely matches that found for words in dozens of different human language samples" (McCowan, et al., 1999, pp. 413).

In a follow-up paper, McCowan, et al. (2002), used Zipf's law to assess the development of a dolphin's whistle repertoire compared to that of a squirrel monkey chuck call repertoire and human language. To categorize the whistles and chuck calls, the Contour Similarity Technique was utilized again. The results suggested that each species exhibited development from more diversity (a more positive regression slope), to more redundancy (a more negative regression slope), and then back to more diversity. McCowan proposed that this showed the three species shared similar developmental contexts, such as a close bond between mother and infant early on. Ultimately, this work was suggested to be a new application of information theory which could significantly advance the field of animal communication studies.

J.P. Hailman has also used Zipf's law in his study of chick-a-dee calls (Hailman, 1994; Hailman & Ficken, 1986; Hailman, Ficken & Ficken, 1985; Hailman, Ficken & Ficken, 1987). The calls of chick-a-dees are much easier to analyze than dolphin vocalizations because they have been categorized into only four different note-types that have very distinguishable features when presented in spectrographic form (refer to Figure 3, reproduced from Hailman, et al., 1985). The note-types have been categorized as A, B, C, and D specifically because the sequences of the calls tend to come in predictable order. That is, it is very common to see sequences like A-A-B-C-D-D or B-C-C-D, but sequences like B-A-B-C-A or D-C-B-A almost never occur.

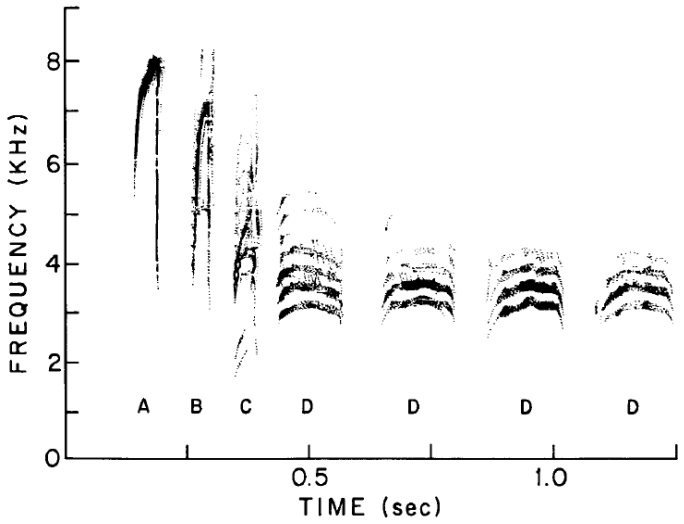



Figure 3. An example spectrogram of a chick-a-dee call sequence.

In 1985, Hailman, Ficken, and Ficken examined the regression slope of frequency versus rank with chick-a-dee calls similarly to McCowan et al.'s work 14 years later with dolphin whistles. Hailman also empirically studied Zipf's ideas of frequency of call versus the length of the call (in number of notes), and the number of call-types/number of calls versus the length (in number of notes). These give rise to what Zipf denoted as economy and variety, which McCowan interpreted from using the regression slope in a plot of frequency versus rank (McCowan, et al., 2002).

In 2005, Suzuki, Buck, and Tyack published a direct rebuttal to McCowan's use of Zipf's law in animal communication analysis which applies to Hailman's use as well. The main argument proposed that tests based on Zipf's law are highly susceptible to false positives, and thus results are un-interpretable. Zipf's law may be a necessary condition for human languages, but it is not sufficient. Suzuki et al. demonstrated this by generating an artificial language from the random process of rolling a die and finding that it had a

similar regression slope to human languages. One must also be skeptical of results comparing animal call-types to human words. As previously noted by Janik (1999), perceptual studies must be conducted to create categories of animal calls that are biologically relevant to the animal. Thus, a great deal of work would be required to determine biologically relevant units for animal calls before comparing them with human words.

*(ii) Entropy*

Since Zipf's law only examines the frequency distribution of a repertoire, McCowan and Hailman both relied on different levels of entropy estimates to analyze the internal organization within their respective repertoires. The zero-order entropy, $E_0$, assumes that call-types are equiprobable, and thus Shannon's equation for entropy reduces to $\log_2 n$. First-order entropy, $E_1$, is based on the actual frequency of occurrence of call-types and is computed as such:

$$E_1 = \sum_i^n p_i \log_2 (1/p_i)$$

Second-order entropy, $E_2$, considers the serial correlation between adjacent call-types based on a matrix of transitional probabilities. Thus, $E_2$ is the weighted average of the component row-entropies calculated from $E_1$, defined as such:

$$E_2 = \sum_{ij}^n p_{ij} \log_2 (1/p_{j|i})$$

where $p_{ij}$ is the joint probability of the $i^{th}$ preceding call-type and the $j^{th}$ succeeding call-type, and $p_{j|i}$ is the conditional probability of $j$ given $i$ as defined in a transition probability matrix. $E_3$ is governed by $E_2$, except $i$ becomes the preceding ordered pairs of call-type. $E_4$ is similar but $i$ is the preceding ordered triplets of call-type. Analysis beyond

$E_4$ is difficult because the amount of data necessary to compute these higher orders grow exponentially. If there are $n$ categories taken $r$ at a time, one needs $n^r$ elements in the data set for a sufficient higher-order entropy approximation (Suzuki, et al., 2005). McCowan summarized these quantities with respect to animal communication as follows: "zero-order entropy measures repertoire diversity, first-order entropy begins to measure simple repertoire internal organizational structure, and higher-order entropies measure the communication system complexity by examining how signals interact within a repertoire at the two-signal sequence level, three-signal sequence level and so forth" (McCowan et al., 1999, pp. 412).

McCowan and Hailman estimated these higher orders of entropy on their respective animal calls and compared them with Shannon's entropy estimates on the English language (Hailman, et al., 1985; McCowan, et al., 1999, 2002; Shannon, 1948). Unfortunately, McCowan was not able to produce interpretable results for estimates greater than $E_1$ due to under-sampling, while Hailman was able to estimate up to $E_4$. Nonetheless, McCowan claimed her analysis showed dolphin whistles shared similar entropic properties as the English language, while Hailman concluded there were noticeable differences between the entropy estimates of chick-a-dee calls and English.

One must hold these results skeptically as it has been previously argued that the comparison of animal call systems to human language is not well defined. Since the entropy estimates rely on call-type units, a valid comparison cannot be made unless we are sure that the respective call-type unit is equivalent to that of an English word.

*(iii) Markov Chains*

In principle, when we observe a sequence of chance events, all of the past outcomes could influence our predictions for the next event. However, this amount of data may be intractable to compute. We can fix this problem if we find a particular correlation between states. Given a finite set of states $S = \{ s_1, s_2, ..., s_q \}$, and a sequence of random variables $X_0, X_1, ...$ , with $X_n = s_i$ denoting a process being in state *i* at time *n*, if $P(X_{n+1} = j \mid X_n=i, X_{n-1}=i_{n-1}, ..., X_0=i_0) = P(X_{n+1}=j \mid X_n=i)$, we say the process follows the Markov assumption and is considered a first-order Markov chain. A first-order Markov chain can be represented by an initial distribution vector $\bar{\boldsymbol{u}} = [ P(s_1), ... , P(s_q) ]$, and a square matrix, **P**, with rows denoted by $X_n = i$ and columns denoting $X_{n+1} = j$. The entry $\mathbf{P}_{ij}$ thus denotes the probability of transitioning from state *i* to state *j*. We have previously referred to this matrix as a transition probability matrix. An example Markov process is weather. There is a good chance that today's weather will have predictive value in considering tomorrow's weather. However, there is a very small chance that the weather from five months ago will have much predictive value for tomorrow's weather. Thus, because tomorrow's weather can be predicted from a finite history of previous day's weather, it can be modeled with a Markov chain.

McCowan et al.'s 1999 study analyzed first-order Markov chains of whistles. Due to under-sampling, not much could be said about the resulting transition probabilities. The data set consisted of the same 186 whistles separated into 27 categories from McCowan & Reiss (1995), which was produced by the Contour Similarity Technique. Thus $27^2 = 729$ whistles would have been needed to properly evaluate the results. No information was given about temporal distances between the whistles studied. Much

more comprehensive data conserving temporal information are necessary in order to run a Markov chain analysis properly.

Hailman et al.'s entropy analysis (1985) included many more calls and found a large drop between $E_1$ and $E_2$, indicating the preceding note in a chick-a-dee call predicts the next note with high accuracy. Hailman and colleagues interpreted this as a characteristic of a semi-Markovian process because a true Markov process would have $E_2 = 0$, rather than the small value they found. Nevertheless, using a first-order Markov chain, Hailman found that an A note is usually followed by another A or by a D, B is followed by itself or by C, C is followed by itself or D, and D is followed by itself or silence.

In a later study, this group of researchers focused on how the call system departed from a first-order Markov process (Hailman, Ficken, & Ficken, 1987). By ignoring the occurrences of repetition in a sequence, Hailman et al. compared the actual distribution of sequences with the expected distribution calculated from the first order Markov chain. Using a $\chi^2$ test where one subtracts the expected frequency from the actual observed frequency, squares the result, and then divides by the expected frequency, one can find the significance in departure from an expected distribution. In particular, Hailman et al. found that sequences of the form (A)(C), where parentheses denote at least one occurrence of a note-type that may have multiple repetitions, were observed far more times than the expected frequency, while (A)(C)(D) sequences were observed far fewer times than the expected frequency. Similarly (A)(B) was observed more than expected whereas (A)(B)(C) occurred less than expected. These results indicate that calls that

begin (A)(C) or (A)(B) are likely to omit the expected (D) notes or (C)(D) notes, respectively.

Hailman et al. also studied the departures from a first-order Markov process as a function of the length of preceding repetitions. For example, when examining the sequence (A)(D), it was found that as the number of A repetitions grew, it was less probable that (D) would follow. Similar results were found for (B)(C), but the most puzzling result for Hailman came from studying (D). As the number of repetitions of D grew, the observed frequency departed significantly from the expected frequency. Also, 86% of all observed chick-a-dee calls contained at least one D-note. The resulting hypothesis to account for the results in this analysis was that D-notes might be flock-identification tags (Hailman, et al., 1987).

*Statistical Analysis of Indus Scripts*:

We have reviewed a number of statistical measures used to analyze dolphin and chick-a-dee vocalizations. Due to insufficient sample size and methodological flaws, it has been difficult to assess a dolphin's acoustic repertoire properly via Markov chains and information theory. If we aim to examine sequences of dolphin vocalizations for combinatorial properties, we need a set of statistics that can reveal strengths of temporal order on an under-sampled data set. In the following study (Yadov, Joglekar, Rao, Mahadevan, & Adhikari, 2009), we find such a framework to analyze a sequence of symbols for evidence of syntax. This framework will integrate previous ideas and introduce new methods of analysis to account for a limited sample size.

*(i) Introduction*

From 7000 BCE to 1500 BCE, a civilization now known as the Indus developed a system of signs which has survived on seals, pottery, and other durable materials. There have been roughly 400 distinct signs identified on over 3000 objects, deemed texts. Refer to Figure 4 for an example Indus text. Due to the scarcity in texts, and a lack of knowledge of any underlying language, the sign system has yet to be deciphered. Most attempts at interpretation have been driven by *a priori* assumptions, resulting in a multitude of varying theories.

Figure 4. An example Indus text. The sign script is located on the top.

Yadov, et al. (2009) approached the signs with a series of statistical measures to find evidence of syntax. Although such an approach cannot yield information on semantics, it makes no *a priori* assumptions. The analysis used a modified corpus of texts void of duplicates, damaged signs, and texts spread over multiple lines since it was unclear how to read such texts. This reduced sample consisted of 417 distinct signs in

1548 texts, with roughly 7000 signs in all, of which no more than 14 signs were found on a single text.

*(ii) Cumulative Frequency Distributions*

The first analysis consisted of plotting the cumulative frequency distribution of signs. It was found that only 69 of the 417 signs accounted for over 80% of the corpus. This follows the distribution previously discussed as Zipf's law since a small number of signs contribute to the majority of the data, while a large number of signs make up the rest. However, this result alone does not directly imply language-like properties as we have previously discussed. To address this question, the researchers plotted the cumulative frequency distribution of text beginners and text enders. Approximately 80 signs that began the texts accounted for 80% of all text beginners while only 23 signs that end texts accounted for 80% of the text enders. Since any of the 417 signs could have begun or ended a text, this is an indication that these text positions are well defined. There is also an implied directionality in the use of signs since there are more beginners than enders accounting for the same percentage.

*(iii) Bi-gram Probabilities*

The main focus of Yadov, et al.'s research was using an $n$-gram model to examine correlations between signs. An $n$-gram model is equivalent to an $(n\text{-}1)^{th}$ order Markov chain, where the likelihood of the next sign in a sequence depends only on the $n$-1 previous signs. Due to the small sample size, only a bi-gram model, or first-order Markov chain, was considered. As previously discussed, a first-order Markov chain is fully specified by an initial sign distribution and a transition probability matrix **P**. Recall that the element $\mathbf{P}_{ij}$ is the probability that sign $i$ transitions into sign $j$. We can also interpret

$\mathbf{P}_{ij}$ as how the observation of sign *i* affects the observation of sign *j* next in sequence. If there is no effect, then $P(s_j \mid s_i) = P(s_j)$, meaning there is no correlation between sign *i* and sign *j*.

Yadov examined matrices of the bi-gram probabilities computed from the corpus, and bi-gram probabilities in the absence of correlation. After assigning a scale of color values to the probabilities, it was clear from just causal observation that the two matrices were very different. Since absence of correlation is expected in a random distribution of signs, the results from this analysis indicate that the signs are not randomly distributed and, thus, there is presence of correlations in the text.

*(iv) Log-Likelihood Significance Test*

Since there was an indication of correlations in the text, a log-likelihood measure of association for bi-grams was computed against the null hypothesis that signs *i* and *j* are independent. The main idea of a log-likelihood test is to measure the statistical significance that a pair of signs occurs together more often than random. Log-likelihood is computed as:

$$2\sum_{ij} O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right)$$

where $O_{ij}$ accounts for observed data and $E_{ij}$ accounts for expected values based on an assumption of independence. The typical method to compute these values is to create a contingency table which displays the frequency counts of how often sign *i* does and does not occur before sign *j* as compared to the rest of the corpus. If the observed values and expected values are comparable, then the log-likelihood is close to 0, indicating the two signs occurred together by chance. However, if the log-likelihood is greater than 0, there

is indication that the signs occurred together significantly more often than expected by chance.

Yadov, et al. computed the log-likelihood test on their data and compared the resulting significant sign pairs with the most frequent sign pairs. It was found that the most frequent sign pairs were not necessarily the most significant.

*(v) Entropy and Mutual Information*

To test further for correlations in a bi-gram model, the information-theoretic measures of entropy and mutual information were computed on the Indus text corpus. Entropy, as previously discussed, can be thought of as the average amount of surprise when observing an event. Thus, entropy is maximized when the probability distribution is random. Comparing a uniform distribution of signs and the real distribution of signs through the computation

$$E_1 = \sum_i^n p_i \log_2(1/p_i)$$

measures the amount of structure present in the texts. A decrease in entropy exhibits less surprise and indicates correlation in the data. Yadov, et al. found the entropy of the corpus distribution was smaller than a random equiprobable sequence of 417 signs.

To measure quantitatively how much the identity of sign *i* reduced uncertainty about the following sign *j*, the mutual information content was computed as follows:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} P(x,y) \log_2 \left( \frac{P(x,y)}{P(x)P(y)} \right)$$

where $P$(x,y) is the joint probability distribution function of X and Y, and $P$(x) and $P$(y) are the marginal probability distribution functions of X and Y, respectively. In the absence of correlation, when X and Y are independent, the joint probability distribution

$P(x,y) = P(x)P(y)$. This results in 0 mutual information since we would be taking the logarithm of $P(x)P(y) / P(x)P(y) = 1$. Mutual information can also be computed as $E_1(X) - E_1(X|Y)$; the entropy of X minus the entropy of X given that we have observed Y. If Y completely determines X, then $I(X;Y) = E_1(X)$. Yadov, et al. found the mutual information content of the Indus corpus was greater than 0, indicating correlation in the data. However, it was also less than the entropy of the corpus, which indicates that the preceding sign *i* does not completely determine sign *j*.

*(vi) n-gram Modeling*

The research team concluded their study by applying their bi-gram model to the restoration of illegible signs in a text. To restore illegible signs, a legible text sequence $S_N = s_1s_2...s_N$ from the corpus was considered and a random sign, $s_x$, was removed. Using the bi-gram model, Yadov, et al. found the most likely choice for sign $s_x$ by evaluating the probability of $S_N$ with all possible choices for $s_x$, and selecting the maximum of these probabilities. For example, the probability for the string $S_3 = s_1s_xs_3$ under the bi-gram model is computed as:

$$P(S_3) = P(<end>|s_3)\ P(s_3|s_x)\ P(s_x|s_1)\ P(s_1|<start>)\ P(<start>),$$

where <start> and <end> are tokens to represent the beginning and end of a text. The bi-gram model was successful in predicting the omitted sign for all test cases considered.

In summary, Yadov, et al. used a series of statistical measures following a bi-gram model on the Indus script corpus to deduce well defined text beginners and text enders, directionality of signs, strong correlations between sign pairs, implications of correlation further back than the preceding sign, and successful prediction of omitted signs in a text. These results reveal the presence of patterned sequences, i.e.,syntax, and

indicate the script can be considered as a formal language. It is still open, however, as to whether the results imply an underlying natural language.

*Conclusion*:

There are a number of issues that must be addressed before sequential analysis can be conducted on dolphin vocalizations, and we have only briefly covered a few of these issues, specifically the categorization of dolphin calls and the estimation of call sequence complexity via information theory. Clearly, there is still much more work to be done in these areas. The literature on categorization of calls has focused primarily on whistles, and even these require more perceptual studies with the dolphins themselves. For example, recent work suggests that dolphins can discriminate among signature whistles within a single category through slight variations in contour (Harley, 2008). Future work must also focus on social use of click trains and perception of burst pulses, vocalizations previously ignored in temporal sequence analysis.

We have also covered various statistical methods that can shed insight into correlations within a sequence of symbols, regardless of whether the symbols are drawn from a set of animal calls or signs from an undeciphered human text. There is a widespread issue of under-sampling in animal communication studies that impacts most quantitative analyses, however, the framework presented by Yadov, et al. (2009) was able to avoid some of the negative impact of a small sample size. Using their selective set of statistical language processing methods, one can infer the presence of syntax in any well defined sign system. The current study defines sequences of dolphin vocalizations in such a way that the Yadov framework can be applied.

To create a suitable analogue to Indus signs, we need to separate the dolphin vocal repertoire into a set of categories. The current chapter has detailed explicitly the difficulty in such a process. Nevertheless, we can use our current understanding of the dolphin's acoustic perceptual system and contextual uses of vocalizations to make a reasonably informed protocol for categorization that serves as a starting point. Although this set of artificial categories is guaranteed *a priori* to be disjoint from a dolphin's perceptions of vocal classes, our ultimate aim is to create a model that approximates sequences of vocalizations a dolphin may produce. The current study initiates this process by developing a method of categorization and examining the sequences in the resulting dataset for elements of syntactic structure via the methods of Yadov, et al.

## Chapter 2: RESEARCH METHODS

*Subjects / Facility*

The focus of this study was temporal sequence analysis of the vocalizations produced by 4 male bottlenose dolphins (aged approximately 8,15,17, and 28 years) residing at a dolphin facility in central Florida. Three hydrophones (Cetacean Research Technology C54XRS, no filter, UltraSound Gate Gain = 4, flat frequency range = 50 kHz) were placed in various configurations of the main tank and back pools where the dolphins were housed. Figure 5 depicts a typical placement of the hydrophones (grey numbered circles). This array allowed us to locate producers of vocalizations in some circumstances.

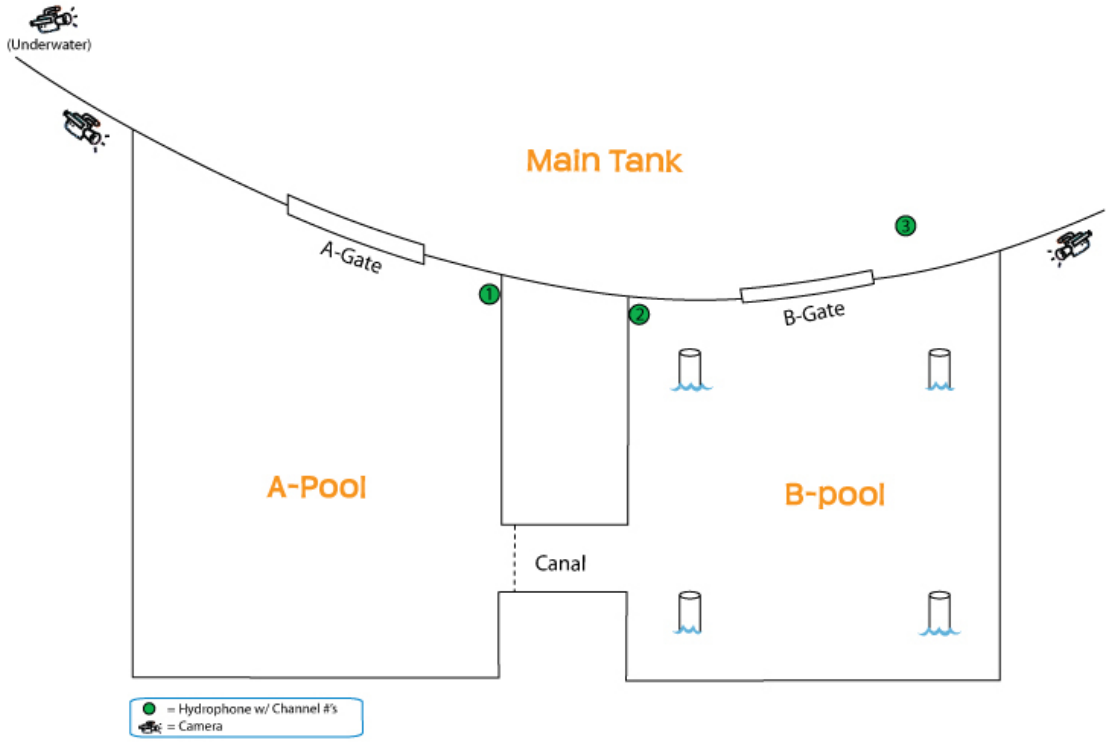



Figure 5. An illustration of the facility housing the dolphins under study.

*Stimulus*

A corpus of vocalizations was derived from recordings of 5-minute intervals spanning different dates and chosen for exploration because they covered a variety of behavioral contexts, e.g., dolphins in isolation, pairs of dolphins vocalizing antiphonally, dolphins in varied groups.

*Procedures*

A lab technician in New College of Florida's Dolphin Lab used a sound analysis program, Avisoft SASLab, to examine spectrograms of the recordings and extract individual vocalizations following a certain protocol (refer to Appendix I for the exact coding procedures used in the current study). This protocol defined how to categorize vocalizations and encode them in a spreadsheet. The spreadsheet was constructed in the following manner (with "-" representing different columns):

> Date - Time - File # - Channel # - Dolphins Present - Vocalization Type (Element/Subevent) - Event # - Element Type (Click/Single Click/Burst Pulse/Whistle) - Label # - Filename of extraction - Duration of vocalization - Start Time - End Time - Classification (Finer Category) - Name of Classifier - Input 1 - ... - Input N

The data in the spreadsheet served as our corpus for the main analysis in this study. Categories of vocalizations could be hierarchical, e.g., sequences, sub-events, and elements. These levels allowed questions concerning information content at different degrees of specificity. We defined a sequence as an event in which any string of dolphin vocalizations occurred within 300ms of each other. We also introduced a category into the corpus labeled "SILENCE" to separate sequences.

*Analysis Procedures*

At the time of our study, the corpus consisted of 1173 labeled vocalizations and 517 SILENCE tokens, representing 516 sequences. Two datasets were constructed from the corpus to investigate the results of using varying categorical levels as input to the sequence analysis program. The first dataset was comprised of only three broad categories of vocalizations (echolocation click trains, brief burst pulses, and narrowband whistles), defined as the highest level of categorization. The dataset considered simultaneous vocalizations as two distinct vocalizations following one another in time. This dataset also preserved the corpus counts, as there were 1173 vocalizations distributed amongst 516 sequences. The second dataset was created from a much finer categorization scheme, resulting in 72 categories of vocalizations. The dataset considered simultaneous vocalizations as a single unit, forming a category with each vocalization identified in the label. This dataset did not preserve the corpus counts due to the handling of simultaneous vocalizations and an omission of isolated echolocation clicks. With this categorization scheme, the dataset consisted of 695 vocalizations distributed amongst 292 sequences. The categories and associated probability distributions of both datasets are included on an accompanying CD. Summary results are available in the next chapter.

The two datasets were used as input to a custom written C computer program that performed a series of statistical tests. The program code is available on an accompanying CD. Drawing from methods used to analyze Indus signs (Yadov, et al., 2009), the program conducted the following analyses on the given dataset input:

1.) Entropy of the frequencies found in the dataset compared to a uniform distribution of categories in the dataset;

2.) Mutual information content found in the dataset compared to an independence assumption between categories in the dataset;

3.) A plot of the frequency distribution of sequence lengths based on number of units for the given dataset;

4.) A plot of the cumulative frequency distribution of all categories, categories that began a sequence, and categories that ended a sequence;

5.) Conditional probabilities of categories found in the dataset compared to conditional probabilities of categories assuming independence;

6.) A log-likelihood measure of association to determine which pairs occurred more frequently than expected.

# Chapter 3: RESULTS

*(i) What level of categorization is suitable for finding patterns?*

The first analysis was designed to determine the level of sub-categorization that best fit the natural tendencies of the dolphin and the question at hand. The categorization scheme of the first dataset was based on a gross acoustic analysis that led to the identification of three categories (echolocation click trains, brief burst pulses, and narrowband whistles). This analysis did not account for simultaneous production of sounds, but instead required extraction of sounds identified in these categories to be considered as if they had been produced alone. Thus, sounds that had overlapped would now follow one another.

Since the investigation's main focus was patterns in vocalizations, this particular analysis began with a comparison of entropy and mutual information based on the probability distribution of the dataset compared to a uniform distribution of categories in the dataset. See Table 1 for the corresponding entropy and mutual information estimates. The estimated entropy of the dataset was 1.9533 bits, similar to the expected entropy of a uniform distribution of four elements (the 3 categories and silence) which was 2.0000 bits ($E_1 = \log_2(4)$). Similarly, mutual information in the data set was low (0.2186 bits), thereby suggesting that knowing the identity of a vocalization type gave almost no information about the vocalization type to come.

| Measure | Random | Dataset |
|---|---|---|
| **Entropy ($E_1$)** | 2.0000 | 1.9533 |
| **Mutual Information ($I$)** | 0.0000 | 0.2186 |

Table 1. The entropy and mutual information of the first dataset.

The entropy and mutual information estimates indicated that the highest level of categorization of the vocalizations did not reflect useful sequence information. Therefore, these measures were calculated a second time using a dataset with a more fine-tuned categorization scheme that took into account our current knowledge of the sophistication in the dolphin's acoustic perceptual system, its use of whistle contours, and its own production of simultaneous vocalizations. This analysis of the same parameters with a more finely-tuned categorization scheme provided a very different picture (Table 2).

| Measure | Random | Dataset |
|---|---|---|
| **Entropy ($E_1$)** | 6.1898 | 4.1907 |
| **Mutual Information ($I$)** | 0.0000 | 1.2913 |

Table 2. The entropy and mutual information of the second dataset.

Since the estimated entropy of the dataset (4.1907 bits) was less than random (6.1898 bits) and the mutual information suggested knowing the identity of a vocalization type reduced uncertainty about the vocalization type to come, we used this categorization scheme in our dataset for other analyses.

*(ii) What is the distribution of sequence lengths in the dataset?*

This analysis calculated the frequency count of sequences by the number of units they contained and was designed to determine if the dataset had enough multi-unit sequences to investigate the existence of patterns in sequences. See Figure 6 for the sequence distribution of the dataset. Although most (162) sequences were only a single unit long, roughly 45% (130 / 292) contained two or more vocalizations.

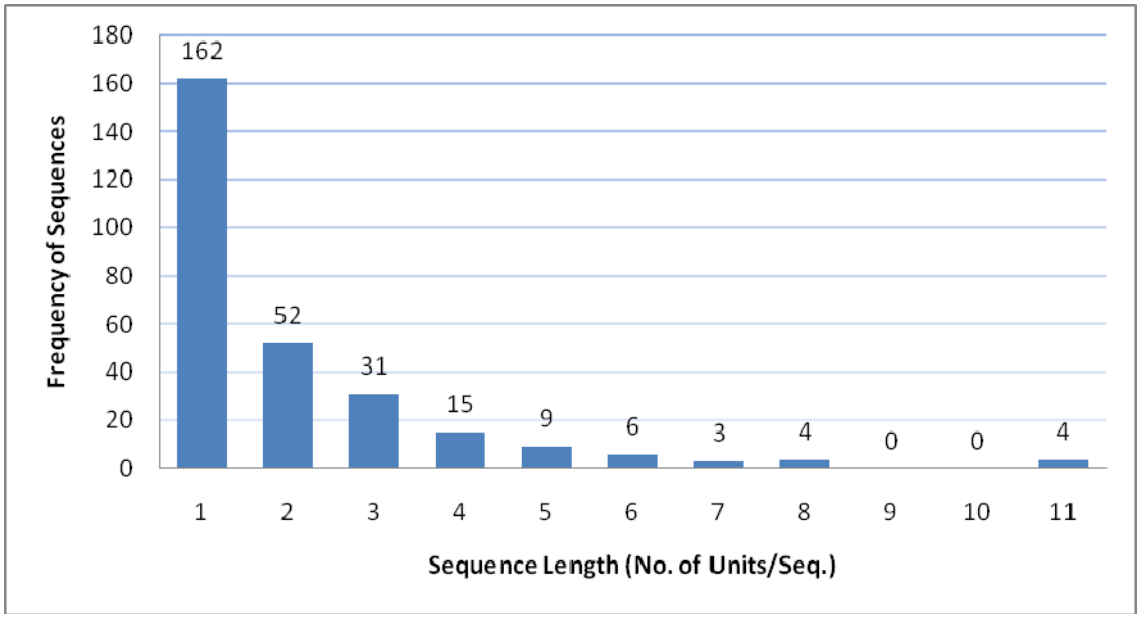



Figure 6. Distribution of sequences by length. Six sequences are not represented in this figure because they were more than 11 units long.

*(iii) How are the categories distributed in the dataset?*

An analysis of the cumulative frequency distribution of vocalization categories revealed a majority of the dataset was comprised of a minority of categories. Subsequent analysis on the cumulative frequency distribution of elements beginning a sequence and elements ending a sequence yielded similar results. See Figure 7 for a plot of these distributions.

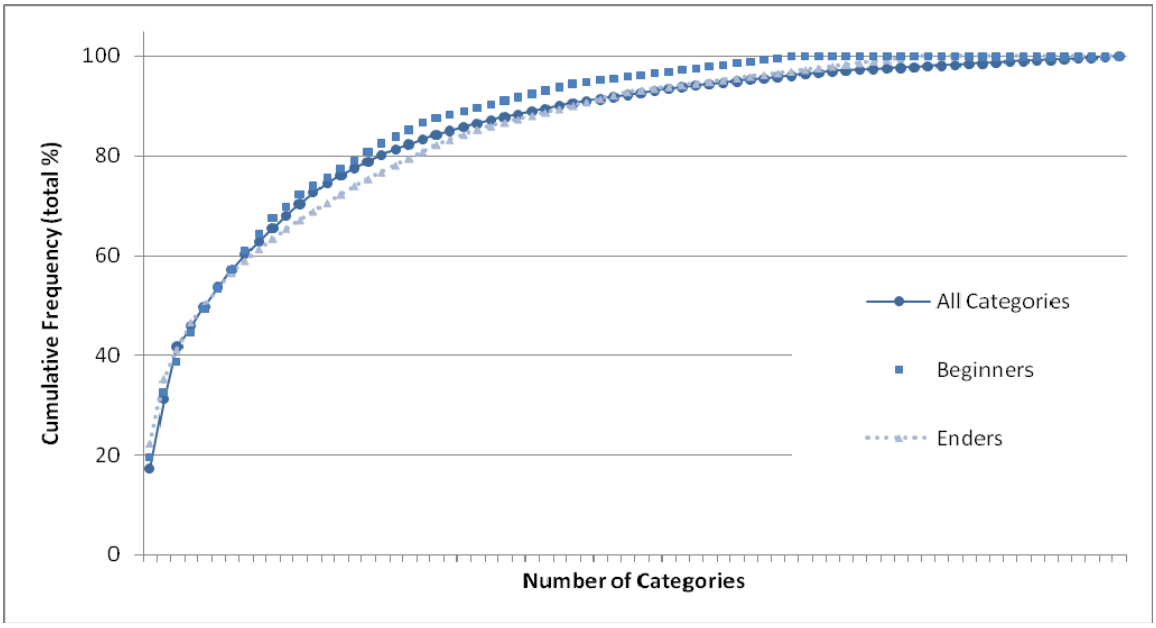



Figure 7. Cumulative frequency distributions of all categories, sequence beginners, and sequence enders.

Eighteen of the 72 categories accounted for over 80% of the vocalizations in the dataset, and 17 of the 72 possible categories that could potentially begin a sequence accounted for over 80% of vocalizations beginning a sequence, with 24 categories never beginning a sequence. Also, 21 of the 72 possible categories that could potentially end a sequence accounted for over 80% of vocalizations ending a sequence, with 15 categories never found at the end of a sequence.

*(iv) What is the strength of correlation between consecutive vocalizations?*

Because entropy and mutual information estimates indicated the presence of structure in the sequences, a log-likelihood measure of association was computed to determine where the structure lay. Table 3 displays the 10 most frequent vocalization pairs observed in the dataset as well as the 10 pairs that had the highest log-likelihood values.

| Vocalization Pair | Frequency |
|---|---|
| <ultra_whistle:SILENCE> | 65 |
| <SILENCE:ultra_whistle> | 57 |
| <SILENCE:burst_pulse> | 38 |
| <burst_pulse:SILENCE> | 38 |
| <burst_pulse:burst_pulse> | 33 |
| <ultra_whistle:ultra_whistle> | 30 |
| <ultra_whistle_inclick:ultra_whistle_inclick> | 30 |
| <SILENCE:upsweep_whistle> | 18 |
| <SILENCE:ultra_whistle_inclick> | 17 |
| <ultra_whistle_inclick:SILENCE> | 17 |

| Significant Vocalization Pairs | Log Likelihood Value |
|---|---|
| <ultra_whistle_inclick:ultra_whistle_inclick> | 74.990518 |
| <ultra_whistle_withBP:ultra_whistle_withBP> | 51.405678 |
| <generic_burst_pulse:generic_burst_pulse> | 49.600525 |
| <ultra_whistle:SILENCE> | 36.324937 |
| <khyber_whistle_inclick:khyber_whistle_inclick> | 31.459185 |
| <SILENCE:khyber_whistle_withBP> | 22.120467 |
| <SILENCE:ultra_whistle> | 19.585574 |
| <SILENCE:khyber_whistle> | 18.213307 |
| <ultra_whistle:ultra_whistle> | 17.485079 |
| < burst_pulse_inclick: burst_pulse_inclick> | 16.357932 |

Table 3. The 10 most frequent vocalization pairs and 10 most statistically significant.

We also examined how the transition probability matrix of the dataset compared to a transition probability matrix assuming conditional independence in the dataset. These probabilities are presented in Figure 8, with the top matrix depicting the transitional probabilities in the absence of correlation. The matrix element [ *a* , *b* ] is a probability determining how likely vocalization *b* will occur after vocalization *a*. The matrices are oriented with the element [0,0] on the bottom left corner and darker shades represent higher probabilities. If consecutive vocalizations were deficient in correlation, we would expect the bottom matrix of Figure 8 to have similar horizontal bands as the top matrix. However, visual inspection reveals the matrices have significant differences. The differences indicate consecutive vocalizations are correlated.

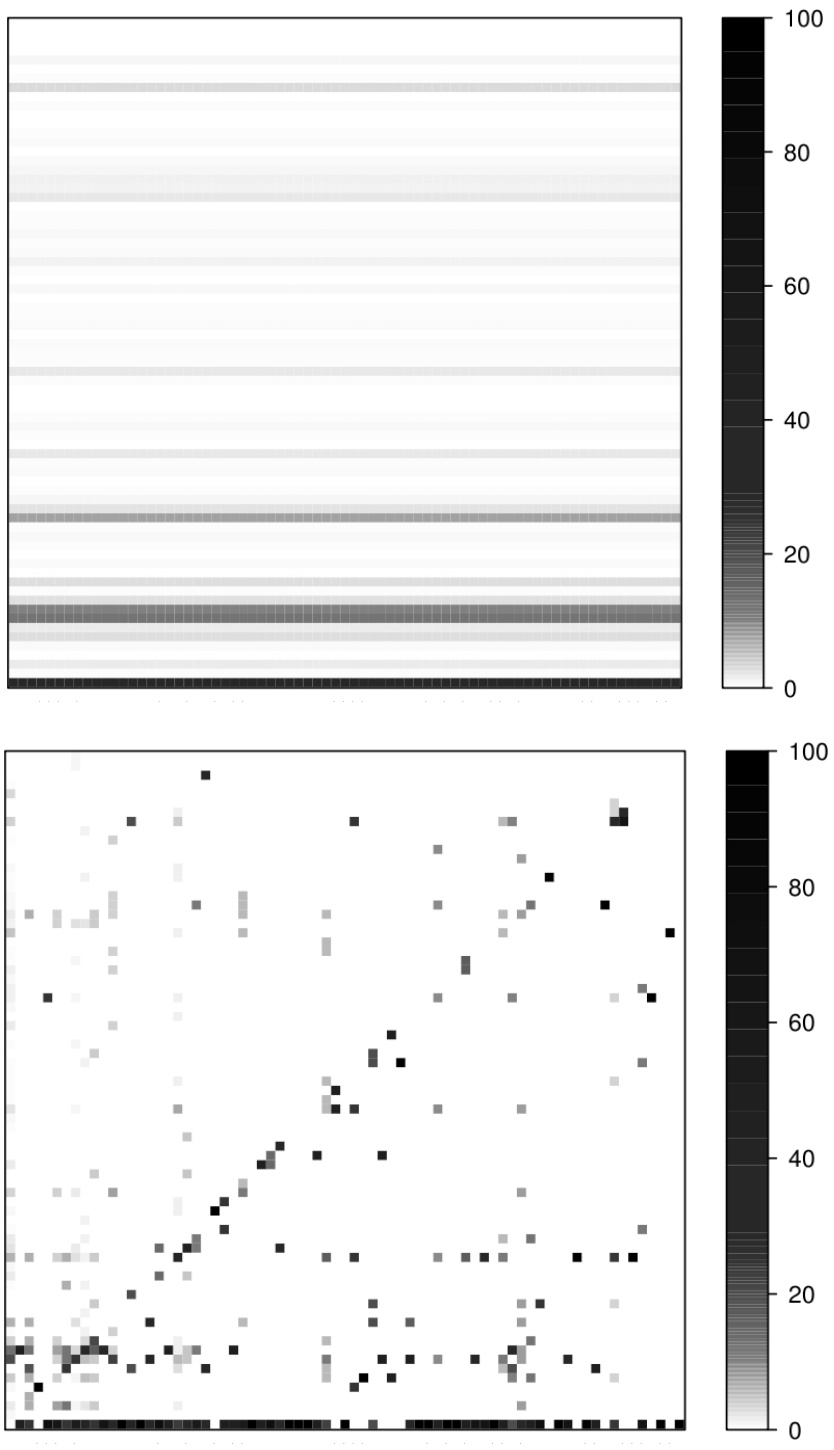



Figure 8. Transition probabilities from the dataset. The top matrix represents an absence of correlation. The bottom matrix represents the distribution of the dataset.

# Chapter 4: GENERAL DISCUSSION

## *(i) Interpretation of Analyses*

The focus of our investigation was to analyze sequences of dolphin vocalizations for evidence of structure. We selected a dataset we believed suitable for finding patterns by considering two categorization schemes and using the information theoretic measures of entropy and mutual information to indicate the internal complexity of the data. Using only three broadly defined categories to model our corpus of dolphin vocalizations resulted in a near uniform distribution with independence between successive units. The use of a more finely tuned categorization scheme that exploited the sophistication of the dolphin's perceptual system provided much more information.

An examination of the frequency distribution of sequences revealed that dolphins produce many multi-unit sequences based on our definition in which 300ms breaks indicated the boundaries of a sequence. These sequences began and ended with well-defined sets of vocalizations with sequence beginners being more strictly defined than sequence enders. Clearly, the sequences have structure.

An entropy and mutual information analysis of the data indicated the presence of correlations between consecutive vocalizations. Analyses comparing observed bi-gram data and expected bi-gram data under random assumptions revealed many differences, further indicating the presence of correlations between consecutive vocalizations. One unexpected result was the frequent occurrence of an "ultra whistle" unit in both tables. This is a category for whistles less than 15ms in duration. Surprisingly, these very common whistles are typically disregarded in analyses of dolphin vocalizations. Both the frequency of their overall occurrence as well as their occurrence in pairs suggests that

omitting these whistles leads to a disregard of what is probably a functional element for dolphins.

In summary, our analysis revealed there are well-defined sets of vocalizations which begin and end sequences, and correlation between successive vocalizations is very strong. Overall, these findings establish compelling evidence that there is significant structure in sequences of dolphin vocalizations, i.e., a syntax.

*(ii) Future Research*

There are a few issues with the current analysis we would like to address in future work. The first is sample size. Under-sampling is an inevitable problem with sequence analysis as the sample sizes necessary for higher orders of approximation grow exponentially with each successive order. Our current corpus of 1173 vocalizations can be modeled sufficiently by a first-order Markov chain if the number of categories produced by the chosen categorization scheme is less than 35. In contrast, the dataset used in the analysis had 72 categories. Our corpus size grows almost daily as we analyze more recordings and extract more vocalizations. With a sufficiently sized corpus, we ultimately aim to use *n*-gram analysis on sequences of dolphin vocalizations in a similar manner as the analyses of Indus script by Yadov, et al.

Another concern is proper categorization of simultaneous vocalizations. The current methods either "flatten" overlapping vocals resulting in the creation of a new transition, or select which category we think is "underneath" and append a label to it indicating the vocalization "on top", resulting in a new category representing both vocalizations. Both methods add more categories and, in general, make the dataset sparser. We hope to find a clever solution around this problem.

There are also a number of ideas we would like to explore in future work that were too disjoint from the focus of the current study to be included here. It would be interesting to see if our bi-gram model could recover ineligible vocalizations within a sequence through prediction, much like the omitted signs in Indus texts (Yadov, et al., 2009). This could help vocalization extraction by generating the most likely category the vocalization should be labeled as due to the events before and after it, and computed from probabilities found in the corpus.

*(iii) Conclusion*

The goal of this study was to examine structure in the sequences of dolphin vocalizations. Since very little has been done to address this question in previous studies, let alone incorporating the entire dolphin repertoire in general communication studies, we had to essentially build a framework of methods from the ground up to develop insight into the problem. The methods have hopefully been clearly outlined for future studies.

We considered the problem conceptually as a stochastic process in which a dolphin produced a discrete sequence of symbols governed by a set of probabilities. We developed a protocol for extracting vocalizations and preserving the temporal order. The extracted vocalizations could then be translated into a finite set of discrete symbols via a categorization scheme. This process generates a set of probabilities that governs the behavior of our model. The extraction protocol and categorization scheme determine a function that takes real dolphin vocalizations over time as input and emits sequences of symbols that we analyzed through statistical language processing methods. We believe we have satisfied the goal of the study under this framework.

REFERENCES


Au, W.W. (1993). *Sonar of Dolphins*. New York: Springer-Verlag.

Au, W.W.L., Popper, A. N., & Fay, R.R. (2000). *Hearing by Whales and Dolphins*. New York: Springer-Verlag.

Brown, J. C., Hodgins-Davis, A., & Miller, P. J. O. (2006). Classification of vocalizations of killer whales using dynamic time warping. *Journal of the Acoustical Society of America, 119*, EL34–EL40.

Brown, J. C., & Miller, P. J. O. (2007). Automatic classification of killer whale vocalizations using dynamic time warping. *Journal of the Acoustical Society of America, 122*, 1201-1207.

Buck, J. R., & Tyack, P. L. (1993). A quantitative measure of similarity for *Tursiops truncatus* signature whistles. *Journal of the Acoustical Society of America, 94*, 2497–2506.

Caldwell, M. C. & Caldwell, D. K**.** (1965). Individualized whistle contours in bottlenose dolphins (*Tursiops truncatus*). *Science*, *207*, 434–435.

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

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

Deecke, V. B., & Janik, V. M. (2006). Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls. *Journal of the Acoustical Society of America, 119*, 645–653.

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

Ficken, M. S., Hailman, E. D., & Hailman, J. P. (1994). The chick-a-dee call system of the Mexican chickadee. *Condor*, *96*, 70–82.

Fripp, D., (2005). Bubblestream whistles are not representative of a bottlenose dolphin's vocal repertoire. *Marine Mammal Science, 21 (1)*, 29–44.

Hailman, J. P., Ficken, M. S., & Ficken, R. W. (1985). The "chick-a-dee" calls of *Parus atricapillus*: A recombinant system of animal communication compared with written English. *Semiotica*, *56*, 191–224.

Hailman, J. P., & Ficken, M. S. (1986) Combinatorial animal communication with computable syntax: “Chick-a-dee” calling qualifies as “language” by structural linguistics. *Animal Behaviour, 34*, 1899–1901.

Hailman, J. P., Ficken, M. S., & Ficken, R. W. (1987). Constraints on the structure of combinatorial “chick-a-dee” calls. *Ethology, 75*, 62–80.

Hailman, J. P. (1994). Constrained permutation in “chick-a-dee”-like calls of the black-lored tit, *Parus xanthogenys*. *Bioacoustics*, *6*, 33–50.

Harley, H. E., (2008). Whistle discrimination and categorization by the Atlantic bottlenose dolphin (*Tursiops truncatus*): A review of the signature whistle framework and a perceptual test. *Behavioural Processes 77*, 243-268.

Harley, H. E. & DeLong, C. M. (2008). Echoic object recognition by the bottlenose dolphin. *Comparative Cognition and Behavior Reviews, 3*, 46-65.

Herman, L. M. & Tavolga, W. N. (1980). The communication systems of cetaceans. In L.M. Herman (Eds.), *Cetacean Behavior: Mechanisms and Functions.* New York: Wiley-Interscience.

Herman, L. M., Richards, D. G., & Wolz, J. P. (1984). Comprehension of sentences by bottlenosed dolphins. *Cognition, 16*, 129-219.

Houser, D.S., Helweg, D.A., & Moore, P.W. (1999). Classification of dolphin echolocation clicks by energy and frequency distributions. *Journal of the Acoustical Society of America, 106(3)*, 1579-1585.

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

Janik, V.M. (1999). Pitfalls in the categorization of behavior: A comparison of dolphin whistle classification methods. *Animal Behaviour*, 57(1), 133-143.

Kemeny, J. G., Snell, J. L., & Thompson, G. L., (1974). *Introduction to Finite Mathematics.* Englewood Cliffs, NJ: Prentice-Hall.

Krützen, M., Mann, J., Heithaus, M. R., Connor, R. C., Bejder, L., & Sherwin W. B. (2005) Cultural transmission of tool use in bottlenose dolphins. *PNAS 2005, 102*, 8939-8943

Lammers, M.O., Au, W.W., & Herzing, D.L. (2003). The broadband social acoustic signaling behavior of spinner and spotted dolphins. *Journal of the Acoustical Society of America, 114(3)*, 1639-1650.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press

Marino, L. (1996). What can dolphins tell us about primate evolution? *Evolutionary Anthropology: Issues, News, and Reviews, 5(3),* 81-86.

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., Hanser, S. F., & Doyle, L. R. (1999). Quantitative tools for comparing animal communication systems: Information theory applied to bottlenose dolphin whistle repertoires. *Animal Behaviour*, *57*, 409–419.

McCowan, B., Doyle, L. R. & Hanser, S. F. (2002). Using information theory to assess the diversity, complexity, and development of communicative repertoires. *Journal of Comparative Psychology, 116*, 166–172.

McCowan, B., Doyle, L. R., Jenkins J.M. & Hanser, S. F. (2005). The appropriate use of Zipf's law in animal communication studies. *Animal Behaviour*, *69*, F1-F7.

McCowan, B. & Reiss, D. (1995). Quantitative comparison of whistle repertoires from captive adult bottlenose dolphins (*Delphindae Tursiops truncatus*): A re-evaluation of the signature whistle hypothesis. *Ethology*, *100*, 193–209.

McCowan, B. & Reiss, D. (2001). The fallacy of 'signature whistles' in bottlenose dolphins: A comparative perspective of 'signature information' in animal vocalizations. *Animal Behaviour, 62*, 1151-1162.

Overstrom, N.A. (1983). Association between burst-pulse sounds and aggressive behavior in captive Atlantic bottlenosed dolphins (*Tursiops truncatus*). *Zoo Biology, 2*, 93-103.

Pierce, J. R. (1980). *An Introduction to Information Theory: Symbols, Signals and Noise*. Toronto, Ontario, Canada: General Publishing.

Richards, D. G. (1986). Dolphin Vocal Mimicry and Vocal Object Learning. In R.J. Schusterman, J.A. Thomas, & F.G. Woods (Eds.), *Dolphin Cognition and behavior: A comparative approach.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Roitblat, H.L., Moore, P.W.B., Nachtigall, P.E., & Penner, R.H. (1991). Natural dolphin echo recognition using an integrator gateway network. *Advances in Neural Information Processing Systems, 3*, 273-281.

Roman, S. (1997). *Introduction to Coding and Information Theory*. New York: Springer-Verlag.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions: Acoustics, Speech, Signal Processing, ASSP-26*, 43–49.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Smith, S. T. (1972). Communication and other social behavior in *Parus carolinensis*. *Publications of the Nuttall Ornithological Club, No. 11*.

Suzuki, R., Tyack, P. L. & Buck, J. R. (2005). The use of Zipf's law in animal communication analysis. *Animal Behaviour*, *69*, F9–F17.

Suzuki, R., Buck, J. R., & Tyack, P. L. (2006). Information entropy of humpback whale song. *Journal of the Acoustical Society of America*, *119*, 1849–1866.

Thomas, J. A. (1986). Communication in dolphins. In R.J. Schusterman, J.A. Thomas, & F.G. Woods (Eds.), *Dolphin Cognition and behavior: A comparative approach.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Tyack, P. L. (1986). Whistle repertoires of two bottlenose dolphins, (*Tursiops truncatus*): mimicry of signature whistles? *Behavioral Ecology and Sociobiology, 18*, 251-257.

Xitco, M.J., Jr., & Roitblat, H.L. (1996). Object recognition through eavesdropping: Passive echolocation in bottlenose dolphins. *Animal Learning & Behavior, 24*, 355-365.

Yadov, N., Joglekar, H., Rao, R. P. N., Mahadevan, I., & Adhikari, R. (2009). Statistical analysis of the Indus script using *n*-grams. *arXiv:0901.3017v1 [cs.CL]*.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

## Appendix I - Categorization Protocol

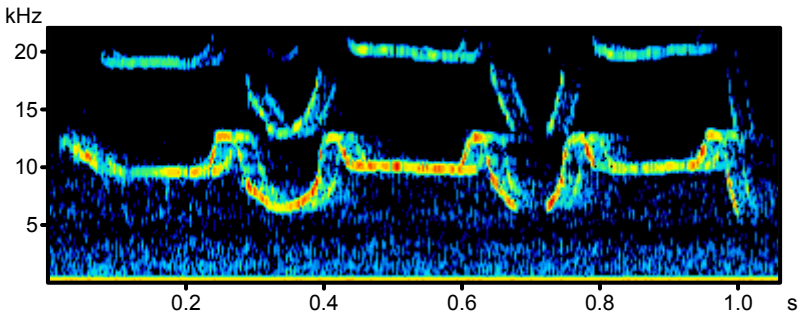
The following protocol was developed specifically to encode vocalizations that may act as phrases in a temporal sequence. There are three different units of sound: events (sequences), sub-events, and elements. Events describe any string of dolphin vocalizations within 300ms of each other. An event might be a whistle and a click train that are separate vocalizations, not at all overlapping, but separated by less than 300ms. Sub-events describe strings of whistles or burst pulses that are close enough to each other to function like a unit, but by previously utilized rules would be considered separate vocalizations. An example might be a multi-loop whistle with an audible break between loops. The threshold currently used for "close enough" is within 200ms. Elements are each separate vocalization.

There are four different kinds of elements: single clicks, click trains, burst pulses, and whistles. Each has specific criteria for which kind of vocalization fits the category. Single clicks refer to a group of less than five clicks, with click being defined as a very short (40-70ms) and loud (180-225 dB re 1 mPa at 1 m) broadband emission (Au, 1993). If there are any other clicks within 300ms of a single click grouping, it is not a single click grouping, but a click train. Click trains are any group of clicks, five or more, that have an inter-click interval of more than 10ms. A click train continues so long as there is another click within 300ms of the last click. Burst pulses are broadband sounds with a very small inter-click interval, or ICI, under 10ms. Burst pulses usually occur in conjunction with clicks, either at the beginning, end or middle of a train. If a click train begins with a burst pulse (or segment where the ICI reaches 10ms or less), the entire section is considered a burst pulse. If the click train ends with a burst pulse, it is labeled

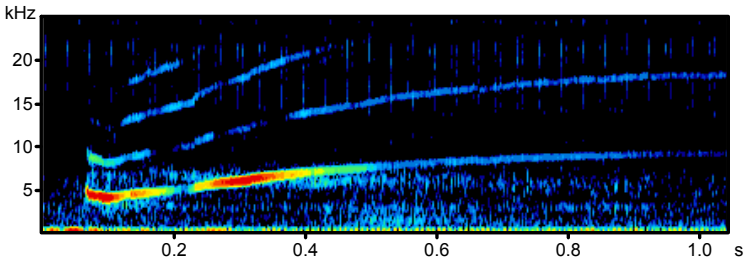
as a click train because the burst pulse is considered a terminal buzz. The only exception is if the ICI is relatively constant throughout the entire click train and the burst pulse at the end is quite abrupt. In that case, the click train is called a burst pulse and the clicks are understood to be a lead-in. If a burst pulse sound occurs in the middle of a click train, the vocalization is categorized as a click train with a burst pulse in the middle of it, two separate elements. For burst pulses, any break in sound necessitates the creation of a new element. Whistles are narrowband sounds, usually less than a second long. Any break in a whistle whatsoever constitutes a new element.

Once this protocol was executed, another pass was made to analyze further the extracted burst pulses and whistles to create a set of finer categories. Burst pulses occasionally were found in short successive numbers of pulses, and sometimes there were distinct bands of energy in a given burst pulse. Inside the broad category of burst pulses, we thus also labeled the number of pulses and the number of bands found within a burst pulse element. Whistles were placed into finer categories based on similarities of frequency contour. The following spectrograms are some exemplars of these whistle sub-categories:

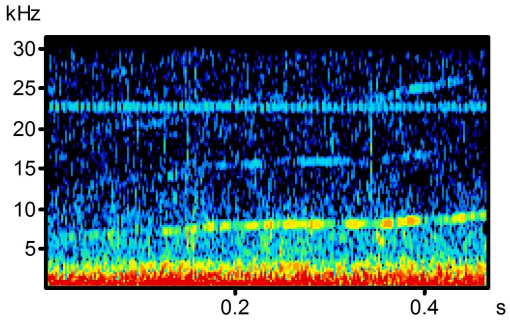
<u>Khyber Signature Whistle</u>



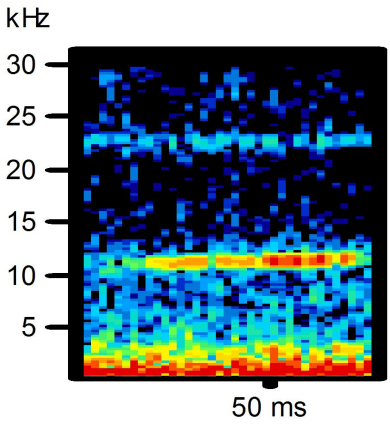
Calvin Signature Whistle




Flat Whistle




Ultra- Flat Whistle




For a complete list of exemplar spectrograms used for vocalization categories, please refer to the included CD.