Information entropy of humpback whale songs[a)]

Ryuji Suzuki[b)]
Department of Electrical and Computer Engineering, University of Massachusetts Dartmouth, North Dartmouth, Massachusetts 02747-2300

John R. Buck[c)]
Department of Electrical and Computer Engineering and School for Marine Science and Technology, University of Massachusetts Dartmouth, North Dartmouth, Massachusetts 02747-2300

Peter L. Tyack
Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543

(Received 2 December 2005; accepted 6 December 2005)

The structure of humpback whale (*Megaptera novaeangliae*) songs was examined using information theory techniques. The song is an ordered sequence of individual sound elements separated by gaps of silence. Song samples were converted into sequences of discrete symbols by both human and automated classifiers. This paper analyzes the song structure in these symbol sequences using information entropy estimators and autocorrelation estimators. Both parametric and nonparametric entropy estimators are applied to the symbol sequences representing the songs. The results provide quantitative evidence consistent with the hierarchical structure proposed for these songs by Payne and McVay [Science **173**, 587–597 (1971)]. Specifically, this analysis demonstrates that: (1) There is a strong structural constraint, or syntax, in the generation of the songs, and (2) the structural constraints exhibit periodicities with periods of 6–8 and 180–400 units. This implies that no empirical Markov model is capable of representing the songs' structure. The results are robust to the choice of either human or automated song-to-symbol classifiers. In addition, the entropy estimates indicate that the maximum amount of information that could be communicated by the sequence of sounds made is less than 1 bit per second.
[DOI: 10.1121/1.2161827]

PACS number(s): 43.80.Ka [WWA] Pages: 1849–1866

I. INTRODUCTION

Payne and McVay (1971) first analyzed the structure of the songs of humpback whales (*Megaptera novaeangliae*). Their analysis relied on human judgments both to determine whether one sound element was identical to another and to determine whether the song fit the hypothesized hierarchical song structure. Subsequent studies (Guinee and Payne, 1988; Miller *et al.*, 2000; Noad *et al.*, 2000; Payne *et al.*, 1983) also relied on human observers for these evaluations. This paper presents methods to remove these subjective judgments by employing an automated classifier to group the units and applying objective information theory techniques to study the song structure.

The rest of this section reviews the previously proposed structure for humpback whale songs and provides an overview of the elements of information theory necessary for this research. Recognizing that information theory falls outside the purview of most animal bioacousticians, the most important theorems and concepts are interpreted in the context of animal communications studies.

A. Songs of humpback whales

Most humpback whales have an annual migratory pattern, breeding in subtropical latitudes during winter, and migrating to high latitude waters to feed in the summer. The vocalizations produced by humpback whales during these feeding and breeding seasons differ greatly. The feeding calls involve a few simple sound patterns produced in simple sequences (D'Vincent *et al.*, 1985), whereas whales produce complex songs during the breeding season. The term *song* is used in animals, such as songbirds and whales, to describe an acoustic signal that involves a wide variety of sounds repeated in a specific sequence.

Humpback songs consist of a sequence of discrete sound elements, called *units*, that are separated by silence. Each song contains a complicated series of more than 12 different units. These units cover a wide frequency range (30–3000 Hz), and consist of both modulated tones and pulse trains. Payne and McVay (1971) proposed a hierarchical structure for humpback song. A song is a sequence of *themes*, where a theme consists of a *phrase*, or very similar phrases, repeated several times. A phrase is a sequence of several units. The song is repeated many times with considerable accuracy to make a *song session*. The reported range of song duration is from 7 to 30 min (Payne and McVay, 1971) or from 6 to 35 min (Winn and Winn, 1978). Winn and Winn (1978) also reported the maximum duration of

[a)]Parts of this paper were presented at the ASA/EAA/DEGA meeting held in Berlin, Germany in March 1999.
[b)]Current address: Speech and Hearing Bioscience and Technology, Harvard-MIT Division of Health Science and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139.
[c)]Electronic address: johnbuck@ieee.org

observed song sessions to be 22 h. Throughout this paper, we use song length to indicate the number of units in a song, and song duration to indicate the number of minutes the song lasts.

All whales in a population are singing the same or very similar songs at a given time, although whales within hearing range do not coordinate to sing the same part of the song at the same time. The songs within a population gradually change over time, so that after several singing seasons few elements of the song have been preserved (Payne *et al.*, 1983). Several reviewers believe that the speed and pervasiveness of this change indicates that singing whales must learn each sound unit and the sequence order that make up a full song (Janik and Slater, 1997; Tyack and Sayigh, 1997). Guinee and Payne (1988) suggested that this song evolution presents a difficult learning and memory task. They proposed that humpback whales increase the redundancy of parts of phrases between adjacent themes as a mnemonic aid, and they found that this redundancy was more common in songs with larger numbers of themes where more material would have to be remembered.

This paper analyzes humpback song with a suite of techniques to determine whether the song contains mathematical properties consistent with hierarchical structure. In this paper, the term *unit* represents the smallest sound element separated by silence, as defined by Payne and McVay (1971), and the term *song* denotes a sequence of units, but the analysis methods developed do not assume the hierarchical syntax proposed by Payne and McVay (1971). A humpback whale song *classifier* is defined as a method of extracting individual units from a continuous sound recording and converting them into discrete symbols (i.e., A, B, C,…) that represent the particular types of sound, preserving the sequence order. In so doing, the classifier absorbs minor variations between sounds by lumping them together as a single symbol. The classifier assigns a new symbol when it encounters a sound element that is deemed sufficiently different from all previously encountered units. Each symbol has no semantics other than representing the particular type of unit. Winn and Winn (1978) performed classification by assigning "approximate phonetic terms" such as "moan," "snore," "cry," and "chirp" to the units. Our analysis of song structure is based on a sequence of units represented by abstract symbols and, therefore, it ignores these acoustic features, duration, and other details of units and gaps between units.

In this paper, the structural constraints in the symbol sequence are referred to as a *syntax* or *grammar*, with no relation to semantics that may or may not exist. This differs from formal language theory where a syntax denotes a set of rules that generates all legal sentences but no illegal ones based on the language's grammar. This paper focuses on the question of which class of syntactic models are the simplest and most accurate for generating sequences of units that match the sequences in humpback songs. Like Payne (1995), no claims positive or negative are made about the semantic content of the songs.

Our analysis addressed three questions. First, what is the quantitative measure of the information conveyed by the songs? Second, can any of the common stochastic models used in animal communication studies reasonably approximate the statistical features of the songs? Third, does a hierarchical model provide the best match to the structure of the songs?

B. Information theory

This section provides an overview of information theory as it pertains to the present study, including several topics not previously employed to analyze animal communications. The primary focus of the section is methods of entropy estimation, but it is first necessary to provide a framework within which these estimates can be properly interpreted. Information theory studies an information *source* $\mathcal{X}$, which produces a sequence, or stream, of discrete symbols. The sequence produced is not assumed to be purposeful or meaningful. Information theory provides techniques to analyze the characteristics of the source, such as the structure of its output, without knowledge of the semantics of the source or its output.

1. Coding bound

At each time i the source $\mathcal{X}$ produces a symbol x_i, for $i=1,2,3,\ldots$. The notation x_i^j represents the subsequence $(x_i, x_{i+1}, \ldots, x_j)$, $i \leqslant j$. An output symbol is called a *letter*, and the set of all possible symbols an *alphabet*, denoted by set $\mathcal{A}$. The entire output sequence is called a *message*, denoted by x_1^∞. Again, there is no assumption that the letters, message, or alphabet are meaningful.

An encoder converts a message to a binary sequence, or *code*, from which a matching decoder can uniquely and exactly reproduce the original message.[1] An encoder is denoted by $f(\cdot)$. Let $\|f(x_1^\ell)\|$ denote the code length of x_1^ℓ, i.e., the number of binary digits, or *bits*, produced by f to encode the message x_1^ℓ of length ℓ.

Definition 1 The *coding rate* of the encoder f for a given sequence x_1^∞ is defined as

$$R_f(x_1^\infty) = \lim_{\ell\to\infty} \sup \frac{\|f(x_1^\ell)\|}{\ell}, \tag{1}$$

that is, the average number of bits per symbol required to code the sequence in the limit of an infinitely long sequence.

There is a lower bound on the coding rate, and there is always some encoder that attains this bound.

Theorem 1 (Coding theorems) For a given message x_1^∞, there is some R' such that

$$R_f(x_1^\infty) \geqslant R' \tag{2}$$

holds for any encoder f. Moreover, there exists an encoder f_{opt} which satisfies Eq. (2) with equality, that is

$$R_{\text{opt}}(x_1^\infty) = R'. \tag{3}$$

Equation (3) is called the coding theorem and Eq. (2) is called the converse coding theorem.

Asymptotically optimal encoders satisfy Eq. (3), achieving the smallest possible code rate in the limit of an infinitely long data sequence. Conversely, encoding x_1^∞ at a rate R less than R' guarantees that x_1^∞ cannot be faithfully recovered from the code $f(x_1^\infty)$, and information will be lost. Conse-

quently, R' is the per letter information contained in the message, and is closely related to the information entropy, as will be shown shortly.

2. Entropy, definition and properties

A source $\mathcal{X}$ may also be considered as a stochastic process, or *process*, producing a sequence of outputs, or a message, controlled by the source's *probability law*. The source's probability law dictates the probability of seeing any given message segment x_i^j for any $1 \leqslant i \leqslant j$ as the output of the source. Let X_i be a random variable representing the source output at time i, X_i^j denote $(X_i, X_{i+1}, \ldots, X_j)$, $i \leqslant j$, and $\mathcal{A}$ an alphabet whose size is $|\mathcal{A}|$. A particular sample output of the random variable is denoted by the lower case letter corresponding to the random variable, e.g., x_i^j. Let $p(x)$ denote $\Pr\{X=x\}$.

Sources where all joint probability distributions are time invariant are called *stationary*. That is, for every possible subsequence $x_i^{i+\ell-1}$ of length $\ell \geqslant 1$, $p(x_i^{i+\ell-1})$ remains the same for all i. Sources where all realizations of the output possess the same statistical properties are called *ergodic*. If the source $\mathcal{X}$ is a stationary ergodic source, it behaves in the same manner in a statistical sense every time the source is started, and also at any given time during the operation. Thus, for a stationary ergodic source, the statistical properties of one long output can be generalized to the properties of the source itself.

A source has memory if the ith output X_i is not statistically independent of all past outputs. If the current output depends *only* on the previous k outputs, the source is a kth-order Markov source.[2] A Markov source where it is possible to reach any state from any state with positive probability in a finite number of steps is said to be irreducible (p. 61, Cover and Thomas, 1991). The Doeblin condition[3] describes a broader class of sources that is less restrictive than Markov sources. The Doeblin condition limits the largest influence that the past may have on the present, while still allowing the influence of the previous outputs from arbitrarily far in the past. Informally, a source satisfying the Doeblin condition blurs the memory of the past output after k symbols; no matter what the output sequence has been up until the current time, all outputs are possible again k symbols into the future. Any irreducible Markov source of a finite order satisfies the Doeblin condition. Also, the Doeblin condition is satisfied if there is an arbitrarily small positive probability of observation error. Thus, almost any practical experimental observation of a stationary source satisfies this condition.

The classes of common information sources can be organized in a list of increasingly strong restrictions, and each class in this list includes all of the subsequent classes as a subset. This list of classes is: (1) Sources which are not stationary nor ergodic, but are governed by a consistent probability law; (2) stationary ergodic sources; (3) stationary ergodic sources that satisfy the Doeblin condition; (4) stationary ergodic irreducible Markov sources of finite order; and (5) independently identically distributed (i.i.d.) sources. Classes appearing later in this list are more restricted in their applicability than earlier ones due to the strong assumptions imposed on the sources in the later class. Note that within the Markov sources, a lower-order Markov source is more restrictive than a higher-order one.

The following discussion of entropy and related source properties summarizes the cumulative contributions of numerous authors. In the interest of brevity, only the strongest results are presented below. The Appendix outlines the progression of these results. Except where noted otherwise, these properties hold for each source class described in the list above.

Definition 2 The *entropy* of a source is defined as

$$H(X_1^\infty) = \lim_{\ell \to \infty} -\frac{1}{\ell} \sum_{x_1^\ell} p(x_1^\ell) \log p(x_1^\ell), \tag{4}$$

where the summation is over every possible subsequence of length ℓ. Throughout this paper, $\log x$ means $\log_2 x$, and $0 \log 0$ is 0. The units of entropy are bits. The entropy of a source is the minimum average number of bits per symbol necessary to encode its messages.

Theorem 2 (Entropy as the coding bound). In the limit, the coding rate of an asymptotically optimal encoder meets the entropy of the source. That is,

$$R_{\text{opt}}(x_1^\infty) = H(X_1^\infty), \quad \text{with probability one.} \tag{5}$$

This theorem allows another interpretation: The entropy of a source represents the average amount of information per symbol that the source transmits in its messages over a long symbol sequence. Given two messages of length ℓ from alphabet $\mathcal{A}$, x_1^ℓ and y_1^ℓ, if the encoding of x_1^ℓ is shorter than that of y_1^ℓ, i.e., $\|f_{\text{opt}}(x_1^\ell)\| < \|f_{\text{opt}}(y_1^\ell)\|$, then x_1^ℓ is said to be more *compressible* than y_1^ℓ. Redundancy in a message can be measured in terms of entropy, and interpreted in terms of compressibility, in light of Theorem 2. A lower entropy of a source implies that the message from the source is more compressible and more redundant. Consequently, entropy is a measure of the average redundancy of the messages from a source.

From the receiver's point of view, entropy is the average measure of *a priori* uncertainty about each successive letter of the source output, and hence the amount of information received equals the amount of uncertainty removed. Equivalently, since a source with little uncertainty is very predictable, entropy decreases with increased predictability.

MacKay (1972) observed "one has no prior reason to regard H as a more biologically significant measure of the *information received* by that organism than, say, the total number or duration of the signals exchanged," (MacKay, 1972, p. 11). However, a properly obtained estimate or upper bound on H is an upper bound of the information received by an organism. Moreover, recent advances in information theory offer new approaches to biological problems beyond the limited methods that MacKay considered.

Entropy is also a measure of the structural constraints and complexity of a source.

Property 1 Additional constraints in the structure of an information source decrease the entropy of that source.

The size of the alphabet restricts the maximum possible entropy to be $H_{\max}=\log|\mathcal{A}|$, which can be attained if and only if each output symbol is independent and uniformly distributed. For a source with entropy $H_{\max}=\log|\mathcal{A}|$, there are $|\mathcal{A}|^{\ell}=2^{\ell H_{\max}}$ possible sequences of length ℓ which may be observed, each with equal probability $|\mathcal{A}|^{-\ell}=2^{-\ell H_{\max}}$. If the output symbols are not independent and uniform, i.e., the source has structural constraints, the entropy H is less than $H_{\max}$. As noted earlier, decreased source entropy implies increased redundancy in the output streams, and this redundancy is quantified as

$$\rho=(H_{\max}-H)/H_{\max}, \tag{6}$$

where $H_{\max}$ is computed for the same alphabet size as the source (Shannon, 1948). Thus, the redundancy ρ is between 0 and 1. Structural constraints imply that some sequences are much more probable than others. The highly probable subset of sequences is called the *entropy-typical set*, and the probability of the other atypical output sequences appearing is very small. This is formally represented by the following theorem.

Theorem 3 (Entropy theorem).

$$\lim_{\ell\to\infty}-\frac{1}{\ell}\log p(x_1^{\ell})=H \quad \text{with probability 1}. \tag{7}$$

Note that this theorem concerns a long individual sequence, and not the average of all sequences. Rearranging Eq. (7) leads to:

Theorem 4 (Typical set). For a positive integer ℓ and $\epsilon>0$, define the entropy-typical set to be

$$\mathcal{T}(\ell,\epsilon)=\{x_1^{\ell}:2^{-\ell(H+\epsilon)}\leqslant p(x_1^{\ell})\leqslant 2^{-\ell(H-\epsilon)}\}. \tag{8}$$

Then, there is a sequence length $\ell'(\epsilon)$ such that for all $\ell>\ell'(\epsilon)$,

$$\Pr\{\mathcal{T}(\ell,\epsilon)\}>1-\epsilon \tag{9}$$

and

$$(1-\epsilon)2^{(H-\epsilon)\ell}\leqslant|\mathcal{T}(\ell,\epsilon)|\leqslant 2^{(H+\epsilon)\ell}. \tag{10}$$

That is, for a small ϵ and sufficiently large ℓ: (1) Almost all of the observed sequences belong to the typical set [Eq. (9)]; (2) all sequences in the typical set are roughly equally likely to occur with probability close to $2^{-H\ell}$ [Eq. (8)]; and (3) the size of the typical set is approximately $2^{H\ell}$ [Eq. (10)]. The atypical set $\mathcal{T}^c(\ell,\epsilon)=\{x_1^{\ell}:x_1^{\ell}\notin\mathcal{T}(\ell,\epsilon)\}$ contains sequences that are improbable or do not occur. When H is significantly smaller than $H_{\max}$, the set of the typical sequences is only a tiny subset of the set of all possible sequences, $\mathcal{A}^{\ell}$, and the size of the atypical set $|\mathcal{T}^c(\ell,\epsilon)|$ is very large. Summarizing this theorem:

Property 2 Stationary ergodic sources with entropy H typically produce $2^{H\ell}$ approximately equiprobable sequences of length ℓ.

Therefore, for a given message length, a source with a larger entropy produces a greater variety of alternative messages than a source with a smaller entropy. Note that this statement concerns only the entropy H, and that the alphabet size $|\mathcal{A}|$ is not a direct factor in the number of typical messages.

Example For the English alphabet of 26 letters plus a space ($|\mathcal{A}|=27$), consider two processes producing very long sequences. The first source is a monkey typing on a keyboard such that each letter is equally probable, and the second source is an English writer. We regard both of these as information sources, and take subsequences of an arbitrarily chosen length 150 from each source output, and call each of them a sentence. The monkey process has the maximum entropy possible for the alphabet size, $H=H_{\max}=\log 27\approx 4.755$ bits, producing all $27^{150}=2^{150\log 27}\approx 5\times 10^{214}$ possible sentences with equal probability. This is the size of the typical set for the "monkey typing" source. Cover and King (1978) estimated the source entropy of English to be approximately 1.3 bits. Thus, English writers typically produce $2^{1.3\times 150}\approx 5\times 10^{58}$ sentences. The size of this typical set is $2^{(\log 27-1.3)150}\approx 10^{156}$ times smaller than the typical set of the monkey process, which produces all possible sentences. The output of the English writer has structural constraints in the form of lexicographical and grammatical rules and the context of the story, limiting the size of the English typical set to be a tiny portion of the size of the monkey typing source typical set.

These interpretations of entropy provide the basis for studying the structure of humpback whale songs. The direct application of Eq. (4) for entropy estimation is difficult for practical problems with a finite set of observations, because the *true* probability law for the sequence of the units in humpback song is not known. Instead, $p(\cdot)$ must be estimated from the available observations, which are necessarily finite in length. One common method is to use $\hat{p}(\cdot)$, the observed *relative frequencies* of events, or the *empirical distribution*. The true distribution and empirical distribution are conceptually different quantities, and without *a priori* knowledge of the source, there is no guarantee that the empirical distribution $\hat{p}(\cdot)$ from an observed output sequence is within a given tolerance of the true probability $p(\cdot)$.

As in any model-based estimation problem, the estimation of the source entropy generally requires: (1) Establishing and justifying a stochastic model for the source, (2) estimating the model parameters (the probability mass function), and then (3) estimating the entropy of the source. Two popular source models used for Step (1) are the i.i.d. model and the empirical Markov model. These models are often assumed under speculation or without justification in the study of animal communications. For example, Beecher (1989) used i.i.d. entropy estimates as a measure of the information embedded in signature calls, implicitly assuming that each call is statistically independent and identically distributed. Gentner and Hulse (1998) used Markov models of varying order with no justification that the sources analyzed fit these models. Another approach to estimation problems is the use of model-free, or nonparametric, methods. In the nonparametric approach, the selection of a model is unnecessary, and therefore the approach is more universally applicable.

Instead of justifying a specific model, we chose to per-

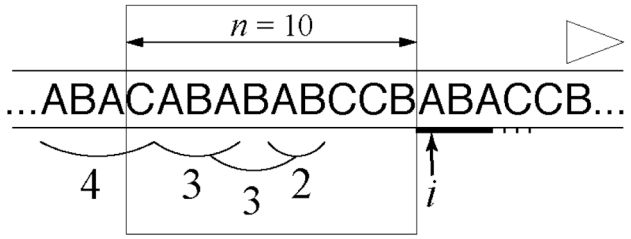

FIG. 1. Example of the sliding window match length. A section of the sample sequence with current index i and window of size $n=10$ are shown. The match length at this index is $L_i(10)=\max\{3,3,2\}=3$. Note that the longer match of ABAC with length 4 is not counted because it begins before the start of the window.

form the entropy estimation with three methods: (1) Estimation with an i.i.d. model; (2) estimation with an empirical Markov model of order 1; and (3) a nonparametric method. Each model's ability to approximate the structure embodied in the humpback songs can be assessed from the resulting entropy estimates using Property 3 (Sec. I B 4).

3. Model-based entropy estimation

The entropy of the i.i.d. and Markov models can be simplified from the general expression in Eq. (4). The model-based estimators result from substituting the observed empirical distributions $\hat{p}(\cdot)$ for the unknown probability mass function (PMF).

Definition 3 (i.i.d. model). The source entropy of an i.i.d. process is estimated from a sample sequence x_1^ℓ of length ℓ by

$$\hat{H}_0 = -\sum_{a\in\hat{\mathcal{A}}} \hat{p}(a)\log \hat{p}(a), \tag{11}$$

where the empirical distribution of the individual letters is

$$\hat{p}(a) = \frac{|\{i{:}x_i = a, i \in [1,\ell]\}|}{\ell}, \tag{12}$$

and the estimated alphabet $\hat{\mathcal{A}}$ is

$$\hat{\mathcal{A}} = \{x_i{:}i \in [1,\ell]\}. \tag{13}$$

That is, the empirical distribution for a letter $\hat{p}(a)$ is the number of appearances of that letter in the sequence divided by the length of the sequence, and the estimated alphabet $\hat{\mathcal{A}}$ is the set of all symbols appearing in the data sequence. For example, for the sequence segment of length $\ell=19$ in Fig. 1, $\hat{p}(\mathrm{A})=7/19$, $\hat{p}(\mathrm{B})=7/19$, and $\hat{p}(\mathrm{C})=5/19$, for the i.i.d. model. The resulting entropy estimate for this unrealistically short example sequence would be $\hat{H}_0 \approx 1.57$ bits. This example is meant for illustrative purposes only, and in practice $\hat{H}_0$ should only be computed for much longer sequences.

The i.i.d. model ignores all statistical dependencies between occurrences of letters by treating them as conditionally independent, and assumes that they are drawn from an identical distribution. The presence of memory in a statistical sense means that the present state or output is at least partially a consequence of the past states. The i.i.d. model fails to represent any effect of memory; hence, it is sometimes called the stationary memoryless model. Such a model is often used as a measure of individual divergence in ecology (Good, 1953; Patil and Taillie, 1982; Peet, 1974). Equation (12) relies on the strong law of large numbers to ensure that the empirical distribution, $\hat{p}(\cdot)$, converges to the true distribution with probability one as $\ell\to\infty$. Therefore, the i.i.d. condition is crucial for this model to be valid, i.e., for $\hat{H}_0$ to converge to the true H.

The simplest source model with memory is the first-order Markov model, where only the immediately previous output influences the current output through the conditional probability law.

Definition 4 (Markov model). The source entropy of a Markov model of order $k=1$ is estimated from a sample sequence by

$$\hat{H}_1 = -\sum_{a_1,a_2\in\hat{\mathcal{A}}} \hat{p}(a_1,a_2)\log \hat{p}(a_2|a_1), \tag{14}$$

$$= -\sum_{a_1,a_2\in\hat{\mathcal{A}}} \hat{p}(a_1,a_2)\log \hat{p}(a_1,a_2) - \hat{H}_0, \tag{15}$$

where the empirical distribution of the conditional and joint probability are obtained by counting the occurrences of each pattern:

$$\hat{p}(a_2|a_1) = \frac{|\{i{:}x_{i-1}=a_1, x_i=a_2, i\in[2,\ell]\}|}{|\{i{:}x_{i-1}=a_1, i\in[2,\ell]\}|}, \tag{16}$$

and

$$\hat{p}(a_1,a_2) = \frac{|\{i{:}x_{i-1}=a_1, x_i=a_2, i\in[2,\ell]\}|}{\ell-1}. \tag{17}$$

Equation (16) is equivalent to counting what fraction of the appearances of a_1 are immediately followed by a_2. For the example in Fig. 1, $\hat{p}(\mathrm{B}|\mathrm{A})=5/7$, $\hat{p}(\mathrm{C}|\mathrm{C})=2/5$, but $\hat{p}(\mathrm{A}|\mathrm{B})=5/6$ since we cannot count the final B because we do not know what symbol follows. Equation (17) is equivalent to counting each possible pair of letters, then dividing by $\ell-1$, the total number of pairs. Referring to Fig. 1, $\hat{p}(\mathrm{A},\mathrm{B})=5/18$, $\hat{p}(\mathrm{B},\mathrm{A})=5/18$, $\hat{p}(\mathrm{A},\mathrm{C})=2/18$, $\hat{p}(\mathrm{C},\mathrm{A})=1/18$, etc…. Substituting appropriate empirical distributions for this unrealistically short segment yields the estimate $\hat{H}_1 \approx 0.98$. Again, this example is only meant for illustration, and in practice $\hat{H}_1$ should only be computed for much longer sequences. More generally, a kth-order Markov model requires estimates of the joint and conditional probabilities from observed blocks of length $k+1$.

Equations (16) and (17) rely on the ergodic theorem to ensure that the empirical joint distribution of the overlapping blocks of length 2 contained in a single length-ℓ block sample sequence converge to the true distribution with probability one as $\ell\to\infty$. It is crucial for the validity of this model that the stationary ergodic and Markov properties be maintained throughout the entire sample sequence. If these assumptions are not valid, the resulting entropy estimates

obtained by using this model may be significantly flawed. Secs. I B 5 and I B 6 discuss additional difficulties in the practical application of Markov model entropy estimates even when all the model assumptions are valid.

4. Model-free entropy estimation

As noted in Sec. I B 1, the encoder f is a data compressor; an encoder compresses the sequence x_1^ℓ to a code of length $\| f(x_1^\ell) \|$ bits, which preserves all of the information in the original sequence and is usually more compact than the original sequence. A *universal* data compressor is an encoder that does not require *a priori* knowledge of the source's probability law (Kolmogorov, 1965). An asymptotically optimal data compressor is one where the average code length approaches entropy as the sequence length increases [Eq. (5)]. Ziv and Lempel (1977) developed the first universal and asymptotically optimal data compression algorithm, called LZ'77 (Wyner and Wyner, 1994; Ziv and Lempel, 1978a). Theoretically, one can use LZ'77 as the encoder $f(\cdot)$ in Eq. (1) with a *long* observed sequence and then consider the resulting code rate $R_f(x_1^\ell)$ as an estimate of the entropy of the source. In practice, this estimator converges to H too slowly in the sequence length ℓ to be useful in most situations. No commonly employed asymptotically optimal universal data compression algorithm converged quickly enough to be useful for the data lengths available in this study.

The nonparametric entropy estimator presented in this section resulted from a performance analysis of the LZ'77 encoder. This estimator provides reliable and robust entropy estimates, and converges to the true entropy H much faster than the LZ'77 approach described in the previous paragraph. Even if the source is nonstationary, or the sample sequence is not long enough to achieve asymptotic convergence, the resulting entropy estimates are an upper bound on the source entropy, and thus is an upper bound on the amount of information transmitted (Theorem 2).

The nonparametric entropy estimate is computed from the *match length* statistics.

Definition 5 (Match length). Match length $L_i(n)$ is defined for a fixed window size n and current sequence index i by

$$L_i(n) = \max\{L : x_i^{i+L-1} = x_{i-k}^{i-k+L-1}, k \in [1,n]\}. \tag{18}$$

Figure 1 illustrates an example of computing the match length. Consider the sequence of symbols shown. The match length is the length of the longest string starting at the present sample x_i that matches a string which begins within the window of n symbols immediately preceding x_i. In Fig. 1, $n=10$, and the longest match $L_i(10)=\max\{3,3,2\}=3$.

Theorem 5 For stationary ergodic sources, the relation between the match length and source entropy is

$$\lim_{n\to\infty} \frac{L_i(n)}{\log n} = \frac{1}{H}, \quad \text{with probability one.} \tag{19}$$

Note that this limit concerns the behavior of the match length at a fixed sequence index i as the window length tends to infinity. A similar result holds for the average match length over the entire sequence.

Theorem 6 For stationary ergodic sources with finite memory,

$$\frac{\mathbf{E}\{L_i(n)\}}{\log n} = \frac{1}{H} + \frac{O(1)}{\log n} \tag{20}$$

for any i and a fixed n. $O(1)$ represents some constant that does not grow with n. A very similar result holds for sources satisfying the Doeblin condition. Thus, for almost any experimental data, this theorem states that the expected match length is roughly inversely proportional to the entropy, with a vanishing constant, $O(1)/\log n$. Replacing the expectation in Eq. (20) with the sample mean over the observed sequence produces the sliding window match length (SWML) entropy estimator.

Definition 6 (SWML entropy estimator). The SWML estimator for the entropy of a stationary ergodic process satisfying the Doeblin condition is

$$\hat{H}_{\mathrm{SW}}(n) = \ell \left[\sum_{i=3}^{\ell} \frac{L_i(n_i')}{\log n_i'} \right]^{-1}, \tag{21}$$

where

$$n_i' = \min\{i-1, n\}. \tag{22}$$

where n is specified window size and n_i' is the effective window size. This estimator is asymptotically unbiased as $n \to \infty$ with probability one. Informally, n_i' is introduced to alleviate transient issues at the start of the data sequence when the index $i < n$.

The SWML estimator has several desirable properties. First, the SWML estimator is applicable to a broader class of sources than the model-based estimators, i.e., Classes (2)–(5) of the list in Sec. I B 2, compared to only Classes (4) and (5) for the Markov model-based estimators, and only Class (5) for the i.i.d. model estimator. Second, the SWML is robust to considerable departures from the assumptions of Eq. (21). When these assumptions are violated, $\hat{H}_{\mathrm{SW}}$ is still a valid upper bound on the entropy. Third, $\hat{H}_{\mathrm{SW}}$ converges rapidly in sequence length ℓ, producing a good entropy estimate from a relatively short data sequence. Fourth, and finally, varying the window size n provides a means to trade off between bias and adaptability to nonstationary sources, as Sec. I B 7 will address.

Despite a superficial similarity, there is no theoretical link between the SWML window size and the Markov model order, or between the SWML estimator and the Markov model. Although the SWML uses a fixed length portion of the immediately previous output, it does not estimate the empirical distributions of Eqs. (16) and (17) for the Markov estimator, Eq. (15).

The ability of the i.i.d. and Markov models to represent the structure of a song can be evaluated by applying each of the three entropy estimators to an observed sequence of units.

Property 3 When comparing two models for a source, the model which most accurately reflects the structural constraints of the source will produce a lower entropy estimate for the source's output.

Consequently, one can use the entropy estimates produced by two models to assess which of the models better fits the structural constraints embodied by a sample sequence. The model producing the lower entropy estimate is a better fit to the constraints of the source. An important caveat to this claim is that care must be taken to ensure that the Markov model estimates are accurate, and not artificially low, as discussed in Sec. I B 5.

Note that the SWML estimator [Eq. (21)] has the least restrictive assumptions of the three methods, and is therefore appropriate for the widest class of sources. Ideally, in the absence of bias, the SWML estimator should yield the smallest entropy estimate of the three estimators. The exception to this statement is if the source is actually an i.i.d. or Markov source. In this case, the SWML entropy estimate should roughly equal the i.i.d. or Markov entropy estimate. If the SWML and i.i.d. or Markov entropy estimates are roughly equal, this indicates that the i.i.d. or Markov model embodies all of the structural constraints apparent in the sample sequence analyzed.

5. Limitations on Markov models

Numerous animal communication studies have employed Markov models, or transition probability analysis (*c.f.*, Gentner and Hulse, 1998; MacKay, 1972; McCowan *et al.*, 1999; Slater, 1973). The empirical application of Markov models has clear limitations on the model order as a function of observed sample length. This section reviews these shortcomings to illustrate the advantages of nonparametric entropy estimation techniques, such as the SWML estimator.

In theory, increasing the order of an empirical Markov model improves its ability to approximate an unknown source. In practice, the finite available observation length ℓ places a limit on the model order. The sequence length required to obtain a reliable estimate of the PMF increases exponentially in model order. Entropy estimates are sensitive to inaccuracies in the PMF. Markov models with improperly high orders will likely not see all of the transitions that the true source produces, or will see them with an incorrect frequency. Such a model is said to *overfit* the data, and generally produces a deceptively low and inaccurate entropy estimate. This is particularly perilous when these erroneously low entropy estimates are interpreted using Property 3 of Sec. I B 4. The low entropy estimates may lead one to conclude incorrectly that the model is a better fit than it actually is.

How long must a sample sequence be to estimate entropy reliably from a Markov model? The data must be sufficiently long to obtain accurate estimates of all the probabilities in Eqs. (12), (16), and (17), or their higher-order analogs. A kth-order Markov model contains $|\mathcal{A}|^{k+1}$ parameters, which are the individual transition probabilities. Consequently, the data observed must exceed $|\mathcal{A}|^{k+1}$ symbols, and preferably $|\mathcal{A}|^{k+2}$. According to Theorem 4, a Markov model of order k and entropy H typically produces roughly $2^{H(k+1)}$ blocks of length $(k+1)$. If the model order k is large enough such that $\ell - k < 2^{H(k+1)}$, some typical blocks do not occur, and therefore the empirical distribution and the entropy estimate are fatally flawed. (Marton and Shields, 1994, 1996). The Markov order k as a function of data length ℓ must grow more slowly than $c\log(\ell)/H$ for the Markov entropy estimate to converge to the true entropy as ℓ goes to infinity (Marton and Shields, 1994, 1996). These results imply that $\ell \geq 2^{H(k+1)}$ is a necessary, though not sufficient, condition for accurate entropy estimation. Lacking prior information about the source, we must assume the most conservative case that $H = H_{\max} = \log|\mathcal{A}|$ and $\ell \geq 2^{(k+1)\log|\mathcal{A}|}$, or $|\mathcal{A}|^{(k+1)}$. This length is an absolute minimum. Preferably, the data length should satisfy $\ell \geq |\mathcal{A}|^{(k+2)}$, to provide accurate entropy estimates for most sources. The $\ell \geq |\mathcal{A}|^{(k+1)}$ requirement may be insufficient unless there is substantial prior evidence that the source's PMF precludes many of the $|\mathcal{A}|^{k+1}$ possible transitions.

The sample length required by these conditions is often onerous. For example, in typical humpback song with $|\mathcal{A}| \approx 20$, even a first-order Markov model requires at least 400–8000 units of song, and a second-order model requires at least 8000–160,000 units. At an average singing rate of 2.5 s/unit, the second-order Markov model requires roughly 5.5 to 110 h of uninterrupted recordings of a single singing whale. For this reason, we limited our Markov model analysis to first order.[4]

The i.i.d. and Markov estimators exhibit a negative bias even when the sample length requirements are satisfied. Section I B 6 presents results on correcting this bias in the entropy estimates. This correction must not be confused with the underestimation of entropy that occurs when Markov models overfit data sequences of insufficient length. Neither bias correction nor bootstrapping approaches can overcome the difficulties of overfitting. Using Property 3 to interpret the low entropy estimates produced by overfitting Markov models results in incorrect conclusions about the appropriateness of higher-order Markov models. Consequently, empirical Markov models have limited viability in entropy estimation. The combined difficulties of the risk of misinterpreting the entropy estimates of overfitted models and the requirement of unrealistically long data sequences argue strongly against the higher-order Markov model analysis of empirical data. One exception to this conclusion involves those rare instances where the source is known *a priori* to be a Markov source, and sufficient data are available.

In light of these strongly worded cautions, how did Shannon (1951) successfully estimate the entropy of printed English? Shannon's insight was to avoid estimating the joint probabilities of the sample data set. Instead, he estimated the error probabilities of human subjects predicting the next letter in an English text from the previous letters. Shannon then used the error probablities to estimate upper and lower bounds on the entropy of English. Subsequent experiments modified the guessing experiment to a gambling game to obtain improved entropy estimates for English (Cover and King, 1978; Miller, 1954; Ch. 6 of Cover and Thomas, 1991; Levitin and Reingold, 1994). Human-based techniques (Cover and King, 1978; Shanon, 1951) produce lower entropy estimates than the SWML estimator (Kontoyiannis, 1997) applied to the same text. Humans thus find less uncertainty in the text than mathematical algorithms. Similarly, it

is possible that humpback whales experience less uncertainty about the sequence of units in a song than the SWML estimator indicates.

Shannon's experiment is often interpreted to support the argument that English can be statistically modeled arbitrarily well by increasing the order of a Markov model. This view was challenged by Chomsky (1956) and Miller and Chomsky (1963). They concluded that: (1) Finite-state Markov models are incapable of representing the recursive hierarchical structures of English grammar; (2) successively higher-order Markov model approximations to English do not converge to true English grammar; and (3) the number of parameters required by higher-order Markov models to reflect grammatical constraints would be immense even if they did model English accurately.

6. Statistical properties of entropy estimators

Interpreting the entropy estimates discussed in Secs. I B 3 and I B 4 requires an understanding of the estimators' statistical properties, such as their bias[5] and confidence bounds. This section assumes that the data sequences are sufficiently long for the model-based methods, precluding the degraded entropy estimates discussed in the previous section. Even when the observed data sequence is sufficiently long, the entropy estimators are negatively biased due to Jensen's Inequality [*c.f.*, Eq. (9) of Wyner *et al.*, 1998; Ch. 2 of Cover and Thomas, 1991].

Theorem 7 (Basharin, 1959). The bias of the i.i.d. estimator for a sample sequence x_1^ℓ from an i.i.d. source is

$$\mathbf{E}\{\hat{H}_0\} - H = -\frac{|\mathcal{A}| - 1}{2\ell} \log e + \frac{O(1)}{\ell^2}. \tag{23}$$

If ℓ is sufficiently large, the term $O(1)/\ell^2$ may be neglected to obtain a bias-corrected estimate.

Definition 7 (bias-corrected i.i.d. estimator).

$$H_0' = \hat{H}_0 + 0.72\frac{|\hat{\mathcal{A}}| - 1}{\ell}. \tag{24}$$

Similarly, the first-order Markov model entropy estimator can be bias corrected as:

Definition 8 (bias-corrected first-order Markov estimator).

$$\hat{H}_1' = \hat{H}_1 + 0.72\frac{|\hat{\mathcal{D}}| - |\hat{\mathcal{A}}|}{\ell}, \tag{25}$$

where the number of all observed units is $|\hat{\mathcal{A}}|$, with $\hat{\mathcal{A}}$ as defined in Eq. (13), and the number of all observed digram transitions is $|\hat{\mathcal{D}}|$. If $|\hat{\mathcal{D}}|$ is less than the true value, the negative bias of $\hat{H}_1$ will be only partially corrected, and there will be some residual bias. Basharin (1959) also provides an expression for the variance of $\hat{H}_0$, but this requires knowledge of the source PMF, so is not applicable to empirical data analysis.

The bias of the SWML estimator cannot be analytically formulated like the biases of the model-based estimators, although it is known to be a positive bias which is $O(1)/\log n$. Thus, the SWML estimator's bias vanishes slowly with increasing window size n. Correcting the bias of the SWML estimator with a bootstrap technique is possible if replicate sample sequences can be generated having the same statistics as the *real* source. For example, Wyner *et al.* (1998) employed a Markov-based bootstrapping bias correction for the SWML estimator when estimating the entropy of DNA. However, in most animal communication research, including the present study, there is no *a priori* justification for assuming that the source follows a Markov model, and therefore it is inappropriate to correct the positive bias of the SWML entropy estimate with Wyner *et al.*'s (1998) technique.

7. Stationarity revisited

As Slater (1973) and MacKay (1972) argued, the processes producing animal vocalizations are not likely to be stationary, which limits the applicability of Markov models. Determining whether an unknown source is nonstationary is not always simple. The autocorrelation technique introduced in Sec. II C provides an indication of the stability of the time-local sequence statistics. If the source is stationary, the correlation values computed for the same lag in different segments of the sequence should be very similar. If large deviations are observed in the correlation estimates, the source is likely to be nonstationary. The statistical properties of many nonstationary sources change so slowly that a short subsequence of the observed data can be regarded as taken from a stationary source; such a source is called *locally stationary*. The i.i.d. and Markov models are strictly limited to stationary sources, and thus produce inaccurate entropy estimates for locally stationary sources. In contrast, the coding bound argument [Eqs. (2), (3), and (5)] justifies the use of the SWML estimator as a practical estimator for nonstationary processes, if the estimate is understood as an upper bound on the source entropy. It is desirable for this upper bound to be as tight as possible. To this end, the SWML estimator employs the following heuristic adaptation for the window size n.

Consider the effect of increasing or decreasing the SWML window size for a globally nonstationary but locally stationary source. Decreasing the window size makes the subsequence x_{i-n}^{i-1} within the window more likely to follow the same statistics as the matching sequence going forward from x_i. Unfortunately, decreasing the window size also increases the positive bias of $\hat{H}_{\mathrm{SW}}$, as can be seen from Eq. (20). Increasing the window size will reduce this bias until the window size n exceeds the limits of local stationarity, and the windowed sequence x_{i-n}^{i-1} no longer has the same statistics as the matching sequence x_i^∞. When the statistics of x_{i-n}^{i-1} and x_i^∞ differ, the observed match lengths are less than they would be for a stationary source of the same entropy. Referring to Eq. (21) reveals that decreasing the match length $L_i(n)$ increases $\hat{H}_{\mathrm{SW}}$. Thus, for a locally stationary source, choosing n to be either too small or too large will increase the positive bias of $\hat{H}_{\mathrm{SW}}$. To balance these competing demands, $\hat{H}_{\mathrm{SW}}$ is chosen to be the smallest entropy estimate obtained over a set of allowable window lengths $\mathcal{I}$, i.e.,

$$\hat{H}_{\mathrm{SW}} = \min_{n \in \mathcal{I}} \hat{H}_{\mathrm{SW}}(n). \tag{26}$$

Thus, the SWML estimator can operate even when the source is not stationary, at the small cost of an increased positive bias.

II. METHODS

The 16 songs analyzed were recorded off the coast of Hawaii from winter 1976 to spring 1978. These recordings of solo whales singing were originally analyzed in Payne *et al.* (1983). Each tape recording began at an arbitrary time during the sequence of units, and finished when the tape ended or the signal faded away. The longest recorded segments in this data set were roughly 45 min, and contained 1000 to 1200 units. Songs shorter than 300 units were rejected because they are insufficiently long for accurate entropy estimates.

The analysis method consists of two steps. First, the audio recordings of the humpback whale songs are converted into a sequence of symbols, where each symbol represents a distinct type of unit. Second, the entropy and correlation properties of the symbol sequences are estimated. Interpreting the entropy estimates' implications for the song structure requires two hypothesis tests. The confidence intervals used for these tests are presented in this section.

A. Classification

Classification necessarily causes an argument: "…there is no guarantee that we will draw the perceptual boundaries in the same place as our study animals" (Tyack, 1998). Ideally, "any human- or computer-generated categorization of vocalizations will need to be validated by testing with the species producing the calls" (Tyack, 1998). Such validating experiments are extremely difficult in the case of humpback songs, because all of the study animals are wild and very large, the units change over time, and the song structure also evolves. Furthermore, the perceptual boundaries of a specific animal may vary over time or depend on the behavioral or communicative context. These boundaries may also vary among the study population. Therefore, one cannot *a priori* assume that there is only one correct classification. The results presented in Sec. III are reassuringly robust to variations between different classifications.

Janik (1999) compared classification methods using human observers with computer-based methods. He remarked that the major disadvantages of human observer classification are bias and lack of reproducibility, but these can be mitigated and minimized by using several observers. Additionally, he noted that the design of automated classifiers must select both appropriate feature parameters and appropriate weightings for these features. In the current study, two human observers and a computer-based classifier were used to supplement each other's shortcomings, and each classification result was analyzed separately. The entire spectrogram of each unit was the input to an unsupervised neural network classifier to avoid the issue of parameter selection.

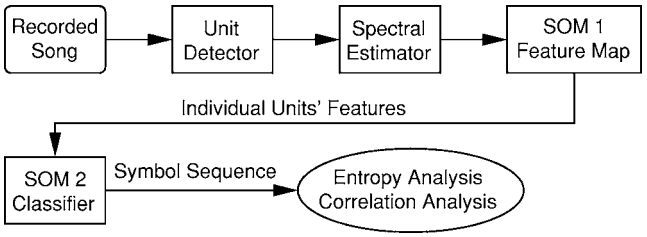

FIG. 2. Block diagram of the automated classifier analysis system for the humpback song.

In an analogous problem, human speech recognition systems, performance improves when the choice of the frequency scale is based on known properties of auditory perception (Jankowski *et al.*, 1995). Although little is specifically known about the hearing of humpback whales, the mammalian ear in general can be modeled as a bank of constant-Q bandpass filters called the cochlea filter (Pitton *et al.*, 1996). Also, humpback whale hearing appears to be low-frequency oriented, roughly matching their vocalization range of 20 Hz to a few kilohertz (Ketten, 1997). For these reasons, constant-Q (logarithmic) frequency-scale spectrograms form the input to the classifiers employed, rather than the more conventional constant bandwidth (linear) frequency-scale spectrogram.

The units in each song were classified using two methods: (1) Manual classification by two human individuals; and (2) automatic classification. Each of the roughly 20–45 min long analog recordings was digitized into a computer using a sampling frequency of 8000 Hz and 16-bit quantization. This digitized signal was used to produce the spectrograms used by both the automated classifier and the humans.

The two human-classified versions were independently produced by the individuals listening to the recorded song with the aid of printed spectrograms. The human classifiers listened to the song as many times as desired, and classified the units by writing the corresponding symbols onto the printed spectrograms. The spectrograms were computed using a polyphase filterbank and covered 30 Hz to 2700 Hz in 80 logarithmically spaced frequency bins and 200 time bins each representing 64 ms. The symbol sequences were typed into a computer file and doublechecked before further processing. The human classifiers were instructed to attach greater importance to the aural qualities of the sound than the printed spectrogram, whereas the automated classification was solely based on the spectrogram. The entropy estimators were applied separately to the symbol sequences produced by each classifier

The rest of this section describes the details of the automated classification. Figure 2 illustrates the automated classification and analysis system in block diagram form. An unpublished nonlinear method (Suzuki, in preparation) detected the signal of the individual units within the recording. The polyphase filterbank computed the spectrograms of these individual units, along with a short sample of the background noise that preceded and followed the unit.

Prior to classification, each unit's spectrogram was preprocessed as follows: (1) The spectrogram was centered within the 12.8 s time window,[6] with any time bins before or

after the signal zero filled; (2) all bins were normalized by the maximum bin; (3) all bins below a threshold were set to zero; (4) soft clipping was applied to all bins above a threshold; and (5) the entire spectrogram was normalized for unit energy. This preprocessing improved the performance of the classification over the range of signal amplitudes and background noise during the song. The processed spectrograms of the units were the inputs to the automated classifier consisting of a sequence of two self-organizing maps (SOMs) (Kohonen, 2001). This classifier assigned a letter to each unit's individual spectrogram, where (ideally) each distinct letter represented a group of similar spectrograms. The sequences of units in the song recording were thus mapped to a symbol sequence. The work of Walker *et al.* (1996) strongly influenced the design of our classifier.

The SOM classifier does not require an *a priori* decision regarding the alphabet size $|\mathcal{A}|$, but only sets the maximum allowable $|\mathcal{A}|$. This is a distinct advantage over classifiers such as the *k*-means clustering algorithm (Linde *et al.*, 1980) or a learning vector quantizer (LVQ) (Kohonen, 2001; Chap. 10 of Duda *et al.*, 2001), since it avoids a major subjective decision about the exact number of classes present in the data. This advantage led us to use the SOM for our classifier in spite of Kohonen's advice advocating LVQs over SOMs for classification and decision tasks (Kohonen, 2001).

The first-stage SOM in the classifier groups units with similar spectrograms, implementing a common SOM topological feature map (Haykin, 1999; Kohonen, 2001). This stage is organized as a relatively large two-dimensional lattice with a large slowly decaying topological neighborhood. The output of this first stage is a matrix of similarity values between the input spectrogram and each neuron's trained synaptic weight. The second-stage SOM takes the similarity matrix for a unit as its input and produces a discrete symbol in $\mathcal{A}$ as its output. The second-stage SOM is organized as a one-dimensional lattice whose length is determined by the maximum allowed value for the alphabet size $|\mathcal{A}|$. This SOM has a small, rapidly decaying neighborhood suitable for classification.

The classifier was independently trained for each song processed. Both SOMs were trained using spectrograms of randomly chosen units from the song. For the first-stage SOM, each unit was equally likely to be chosen as the input. After the first-stage SOM was trained and the weights were fixed, the second-stage SOM was trained twice, using the similarity matrices output by the first SOM for randomly chosen units. For the first training run of the second-stage SOM, each unit spectrogram was equiprobable. For the second training run of the second-stage SOM, each output letter was chosen equiprobably, and then an individual spectrogram for an instance of that letter was chosen equiprobably from all spectrograms classified as that letter during the first run. Therefore, during the second training run, spectrograms belonging to infrequently appearing units were chosen more frequently than the spectrogram of common units, but all possible classifier output letters received approximately the same number of training iterations.

Both SOMs were trained with the standard adaptation techniques found in Kohonen (2001) and Haykin (1999), with typical parameters given in Table I using the notation of Chapter 9 of Haykin (1996). The topological neighborhood function was a Gaussian function as suggested in Haykin (1999). The only modification from the standard SOM neuron update algorithm was that, for the first-stage SOM, the spectrogram matrix could be shifted by up to Δt bins in time and Δf bins in frequency to determine which neuron's weights were the most similar to the current spectrogram. For the first 2000 iterations of training, the neurons' weights were then updated using the unshifted original spectrogram. After 2000 iterations, the neurons' weights were updated using the shifted spectrogram which gave the best fit to the optimal neuron. For the purpose of these shifts, both the spectrogram and the neuron weights were considered to be zero outside their defined index range. Table I also includes typical values for Δt and Δf.

TABLE I. Neural network parameters. These are the typical parameters used for the two stages of self-organizing maps (SOMs) in the automatic classification system.[a]

Parameter	First stage	Second stage
Map size	8×8	1×26
Input dimension	80×200	8×8
Iterations	40,000	80,000
σ_0	8	7
τ_1	850	8
η_0	0.1	0.03
τ_2	1000	800
Δt	10 bins	⋯
Δf	6 bins	⋯

[a]The notation for the SOM parameters in this table follows Chapter 9 of Haykin (1999).

Once the SOMs for both stages of the classifier converged, the spectrograms of the units were processed again to produce the automated classifier's symbol sequence for that song.

B. Information theory analysis

In order to determine whether the i.i.d. and Markov models accurately represented the structure embodied by each symbolized humpback song, the model-based entropy estimates and the SWML entropy estimate were used to test two hypotheses. The first hypothesis is that $H_0 \leqslant H_1$, with the alternate hypothesis being $H_0 > H_1$. The second hypothesis is that $H_1 \leqslant H_{\mathrm{SW}}$, with the alternate hypothesis being $H_1 > H_{\mathrm{SW}}$. Decisions of these tests are interpreted with Property 3 of Sec. I B 4.

The two hypotheses are tested for the significance level of 0.05 using the following bootstrap technique. For the first hypothesis ($H_0 \leqslant H_1$), 1000 independent bootstrap sequences are generated for each song using an i.i.d. source whose PMF is the empirically observed distribution for the song. Each of these bootstrap sequences has the same length as the original song sequence. The bootstrap sequences are used to obtain 1000 first-order Markov entropy estimates. The 50th lowest of these entropy estimates is chosen as the bound on the one-tailed 0.95 confidence interval for the source entropy measured with $\hat{H}_1$, under the assumption that the actual

TABLE II. Entropy estimates from 16 humpback whale songs using the three estimators described in Secs. I B 3 and I B 4. All entropy estimates are in bits. The i.i.d. and Markov values are bias corrected as described in Sec. I B 6. The number in parentheses next to the $\hat{H}_{\mathrm{SW}}$ indicates the window size n selected as described in Sec. I B 7. $\hat{H}_{\mathrm{SW}}$ is consistently less than both $\hat{H}_0$ and $\hat{H}_1$, demonstrating that the Markov and i.i.d. models fail to capture the full structure of the songs. The strength of the song structure is also apparent in the high redundancy values for all songs.

Date (Tape No.)	Length (units)	$\lvert\hat{\mathcal{A}}\rvert$ (units)	$\log\lvert\hat{\mathcal{A}}\rvert$ $H_{\max}$	i.i.d. $\hat{H}_0$	Markov $\hat{H}_1$	SWML $\hat{H}_{\mathrm{SW}}(n)$	Redundancy $\hat{\rho}$
26 Dec 1976 (1A)	840	21	4.39	4.02	0.84	0.58(10)	0.87
04 Jan 1977 (2)	1103	17	4.09	3.79	1.13	0.64(10)	0.84
01 Feb 1977 (1-1)	972	27	4.75	4.13	0.99	0.47(10)	0.90
04 Mar 1977 (1A-2)	978	25	4.64	3.93	0.75	0.33(11)	0.93
10 Mar 1977 (1A-1)	967	27	4.75	4.22	1.17	0.68(11)	0.86
10 Mar 1977 (1A-2)	577	26	4.70	4.03	0.99	0.57(11)	0.88
12 Apr 1977 (4A)	805	33	5.04	4.54	1.10	0.56(10)	0.89
17 May 1977	1021	17	4.09	3.61	1.12	0.36(10)	0.91
01 Feb 1978 (1A)	1101	26	4.70	4.38	1.10	0.56(11)	0.88
05 Feb 1978 (1A)	976	18	4.17	3.77	0.81	0.51(13)	0.88
05 Feb 1978 (1B)	959	18	4.17	3.88	0.83	0.58(11)	0.85
05 Feb 1978 (2A-1)	380	23	4.52	4.37	0.98	0.52(13)	0.88
05 Feb 1978 (2A-2)	488	25	4.64	4.36	1.05	0.64(13)	0.86
05 Feb 1978 (2B-1)	438	22	4.46	4.26	0.94	0.67(15)	0.85
05 Feb 1978 (3A-2)	436	21	4.39	4.37	1.01	0.60(11)	0.87
07 Feb 1978 (1B)	662	26	4.70	4.45	1.10	0.79(15)	0.83

humpback song is generated with an i.i.d. model. If the observed $\hat{H}_1$ for the song is below this value, the null hypothesis is rejected. A similar bootstrap procedure is used for the second hypothesis ($H_1 \leqslant H_{\mathrm{SW}}$), except that the empirical first-order Markov model is used to generate the 1000 replica sequences, and SWML estimator is used for the entropy estimation. The SWML entropy estimates were not constrained to use the same window size n in Eq. (26) which obtained the minimum H_{SW} for the song, but rather could range over the full set $\mathcal{I}$ for each replica sequence. The resulting confidence bound is for the source entropy measured with H_{SW} assuming that the actual humpback source is a first-order Markov source.

C. Correlation analysis

Rejecting both null hypotheses described above indicates that neither the i.i.d. nor the first-order Markov models are adequate to produce the structure of the observed humpback song. This does not rule out the possibility that the song was produced by a higher-order Markov model. As discussed in Section I B 5, entropy estimates for second- (and higher-) order Markov models will be unreliable for humpback songs with an alphabet of size $|\mathcal{A}| \approx 20$ and length $\ell \approx 300-1200$ units. Instead, correlation analysis can be used to reject the possibility of a higher-order Markov model.

The discrete sequence correlation of two symbol sequences x_1^{ℓ} and y_1^{m} at lag λ is defined by

$$r(x_1^{\ell}, y_1^{m}, \lambda) = \frac{\left|\{i : x_i = y_{i+\lambda}, \max(1, 1-\lambda) \leqslant i \leqslant \min(\ell, m-\lambda)\}\right|}{\min(\ell, m, \ell+\lambda, m-\lambda)} \tag{27}$$

for $|\lambda| < \min\{\ell, m\}$. The numerator represents the number of symbols of x_i and $y_{i+\lambda}$ that agree within the overlapping section. The denominator is the length of the overlapping section of the sequences after y has been shifted by λ. Note that this definition can also be used for autocorrelation if y is set equal to x.

A Markov model of any order is stationary, so the value of the short time autocorrelation $r(x_n^{n+L}, x_n^{\ell}, \lambda)$ should be independent of the value of n chosen, i.e., which segment of the sequence is used for the autocorrelation. In practice, the autocorrelation for a fixed value of λ will fluctuate slightly with n, but should be very similar if the process is stationary. If the estimates of the autocorrelation for different values of n are only similar over a range of lags $0 \leqslant \lambda \leqslant v$, but diverge for $\lambda > v$, this implies that the source is nonstationary, but may be considered locally stationary for windows of length v or less. This value of v determines the range of window sizes $\mathcal{I}$ used for the SWML estimator in Eq. (26).

To determine the stationarity of the humpback song, disjoint sections of longer songs was autocorrelated using two 151-unit sections from the song. Specifically, $r(x_{300}^{450}, x_{300}^{\ell}, \lambda)$ and $r(x_{600}^{750}, x_{600}^{\ell}, \lambda)$ were compared to see if the correlation function differed early and late in the recording.

In addition to this short-time autocorrelation analysis, global autocorrelations were computed using $r(x_1^{\ell}, x_1^{\ell}, \lambda)$ for the songs. These autocorrelations, considered as a function of λ, reveal the timescales of the dependencies and periodicities in the symbol sequence. Sequences which are periodic or quasi-periodic with period N will produce global autocorrelations with the same period. Comparing global autocorrelations from different songs provides information about how the periodicity changes between songs.

III. RESULTS

A. Entropy analysis

Table II presents the three entropy estimates for each song, converted to symbol sequences by one of the human

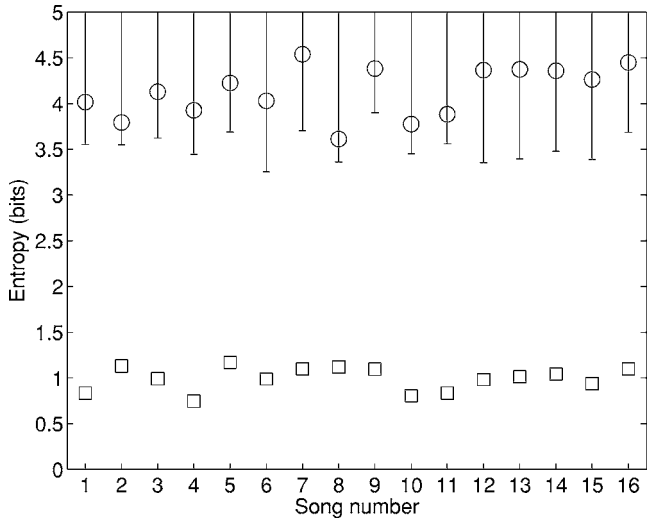

FIG. 3. The values of $\hat{H}_0$ (circles) and $\hat{H}_1$ (squares) plotted for each song in the same order as Table II. The interval plotted for each song indicates the lower limit of the one-tailed 0.95 confidence region for $\hat{H}_1$ under the hypothesis that $H_0 \leqslant H_1$. For all songs, $\hat{H}_1$ lies below the bars, supporting the conclusion that $H_1 < H_0$ with significance $p < 0.05$. This indicates that the humpback songs contain a temporal structure that partially depends on the immediately previous unit within a song, and that the i.i.d. model failed to capture the structures embodied by the humpback songs.

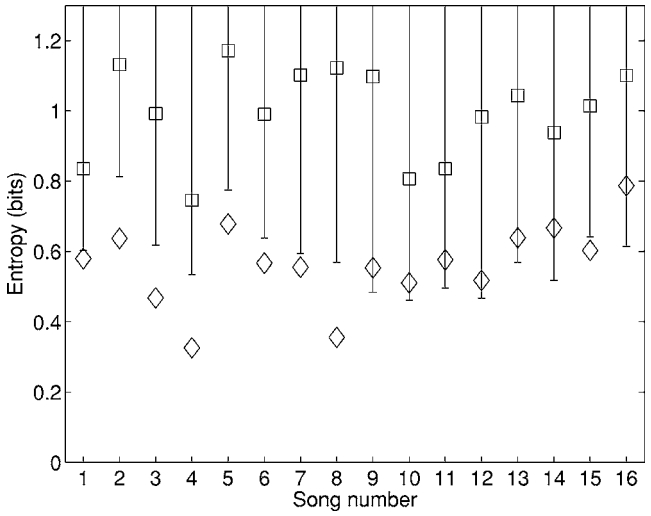

FIG. 4. The values of $\hat{H}_1$ (squares) and $\hat{H}_{\mathrm{SW}}$ (diamonds) plotted for each song in the same order as Table II. The interval plotted for each song indicates the lower limit of the one-tailed 0.95 confidence region for $\hat{H}_{\mathrm{SW}}$ under the hypothesis that $H_1 \leqslant H_{\mathrm{SW}}$. For 9 of the 16 songs, $\hat{H}_{\mathrm{SW}}$ lies below the interval, supporting the conclusion that $H_{\mathrm{SW}} < H_1$ with significance $p < 0.05$. This demonstrates that the Markov model failed to capture all of the structure embodied by the majority of the humpback songs we analyzed, and that the humpback songs contain temporal structure spanning over the range beyond immediately adjacent units.

classifiers. The values given for $\hat{H}_0$ and $\hat{H}_1$ are the bias-corrected values, and the values $\hat{H}_{\mathrm{SW}}$ use window sizes in the range $\mathcal{I}=[10,40]$. The number in parentheses following $\hat{H}_{\mathrm{SW}}$ is the window size yielding this minimum value in Eq. (26). Table II also includes the estimated redundancy $\hat{\rho}$ for each source, using $\hat{H}_{\mathrm{SW}}$ for the entropy H in Eq. (6). The high value of $\hat{\rho}$ for each song indicates that the song structure is strongly constrained, resulting in an entropy much less than $H_{\max}$ for the observed alphabet size $|\mathcal{A}|$.

Figure 3 presents the i.i.d. (circles) and Markov (squares) entropy estimates for individual songs with one-tailed 0.95 confidence bounds for the null hypothesis, $H_0 \leqslant H_1$. Since all $\hat{H}_1$ lie below the confidence bound, we reject the null hypothesis ($p \leqslant 0.05$) and conclude that $H_1 < H_0$. This means that the i.i.d. model failed to capture the structure embedded in all humpback songs analyzed, and also that the humpback songs contain a temporal structure that at least partially depends on the immediately previous unit within a song.

Figure 4 presents Markov (squares) and SWML (diamonds) entropy estimates for individual songs with one-tailed 0.95 confidence bound for the null hypothesis, $H_1 \leqslant H_{\mathrm{SW}}$. 9 out of 16 songs had $\hat{H}_{\mathrm{SW}}$ below this confidence bound, and for these songs we conclude that $H_{\mathrm{SW}} < H_1$ ($p \leqslant 0.05$). For the seven other songs, the test did not conclude that the $\hat{H}_1$ and $\hat{H}_{\mathrm{SW}}$ were sufficiently different at this significance level. We offer two possible explanations to account for this result. All of these seven songs were recorded in a one week period in early February 1978. Since they are likely to be quite similar, those data points may not be statistically independent. Considering only one song in any given month, the second hypothesis test concludes $H_{\mathrm{SW}} < H_1$ for six out of seven months. Furthermore, we note that $\hat{H}_1$ may have residual negative bias (Sec. I B 6) if $|\mathcal{D}|$ was underestimated, and $\hat{H}_{\mathrm{SW}}$ has positive bias (Sec. I B 6). Both sources of bias work to favor the null hypothesis by requiring stronger statistical evidence to reject the null hypothesis, but we disregarded these residual bias considerations in our analysis. Therefore, it is expected that the true test results may be more significant than the results presented in Fig. 4, although we have no way to demonstrate this. Although the result is less clearcut than the one shown in Fig. 3, the conclusion from Fig. 4 is that the Markov model failed to cap-

TABLE III. Automatic versus manual classification. The gaps between the different entropy estimates for each song persist regardless of the classifier used. Consequently, the conclusions about the song structure are robust to the classification method chosen. All entropy estimates are in bits.

		4 Jan 1977			17 May 1977		
		Auto	Human 1	Human 2	Auto	Human 1	Human 2
i.i.d. model	$\hat{H}_0$	2.84	3.71	3.79	4.38	3.61	3.61
Markov model ($k=1$)	$\hat{H}_1$	2.48	1.37	1.13	2.69	1.26	1.12
Sliding window	$\hat{H}_{\mathrm{SW}}(n)$	1.95(10)	0.89(10)	0.64(10)	1.59(18)	0.47(10)	0.36(10)

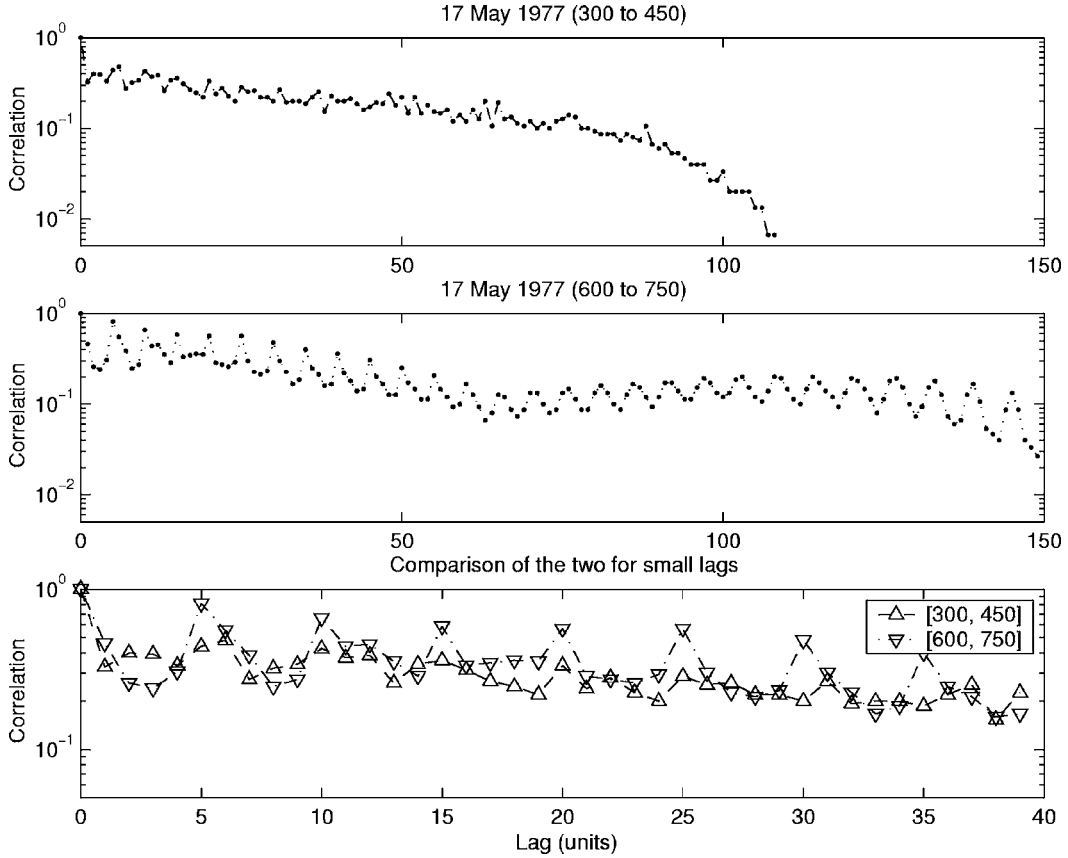

FIG. 5. Two short time autocorrelation estimates for a typical song. The top two panels plot the correlations for units in the ranges [300, 450] and [600, 750], respectively. Note that the correlation functions diverge substantially for lags greater than 100 units, indicating that the song is nonstationary. The third panel shows both correlations on the same axes for small lags. The values generally agree closely to each other with lags smaller than 40, indicating that the song may be considered locally stationary over this range. Both correlations have a strong oscillation with a period of about 6, corresponding to the typical phrase length of Payne *et al.* (1983).

ture all the structure embodied by the majority of songs we analyzed, and also that the humpback songs contain a temporal structure which spans beyond the immediately previous unit within the song.

To control for observer bias, the songs were classified by two humans and the SOM classifier described in Sec. II A. The human classifiers disagree on roughly 5% of the units for these songs, suggesting that there is no clear correct classification for all of the units. The automatic classifier disagrees with the human observers for roughly 10% of the units. Table III contrasts the entropy estimates obtained using each of the three classifiers on two songs. The changes in the entropy estimates between classifiers for a given model reflect the disagreements about the unit classifications. The entropy estimates vary for the different classifiers, but all three estimators exhibit substantial gaps between $\hat{H}_0$ and $\hat{H}_1$ and between $\hat{H}_1$ and $\hat{H}_{\mathrm{SW}}$. These gaps between the entropy estimates are also robust to variations in the SOM parameters. Therefore, the fundamental conclusions above about the songs' structure using Property 3 of Sec. I B 4 are robust to the choice of classifier. We speculate that the strong structural constraints of the songs prevent the perturbations in unit classifications from closing the gaps between the entropy estimates.

B. Correlation analysis

Local autocorrelation functions were computed to test the stationarity of the songs. The results for May 17, 1977 are shown in Fig. 5. The top panel plots $r(x_{300}^{450}, x_{300}^{1021}, \lambda)$ as a function of the lag λ, while the middle panel plots $r(x_{600}^{750}, x_{600}^{1021}, \lambda)$ for the same song. Both plots have logarithmic vertical axes. Comparing the top two panels of Fig. 5 reveals that the correlation curves differ substantially for lags of 100 units or more. This difference between the curves demonstrates that the song statistics are nonstationary. Any irreducible empirical Markov model has stationary statistics, so consequently no such Markov model with an order less than the maximum possible length of the song can capture the structure of the songs. Lastly, note that the oscillations with a period of 6–8 units in the autocorrelation function indicate that there are repetitions in the song with that period.

A possible objection arises to the use of $\hat{H}_{\mathrm{SW}}$ for the humpback songs, since the songs do not have stationary statistics and Eq. (21) assumes a stationary source. We offer two rationales in response to these points. First, even when the assumptions of the SWML entropy estimator are violated, the estimator produces an upper bound on the entropy (Kontoyiannis *et al.*, 1998), so the true entropy is expected to be

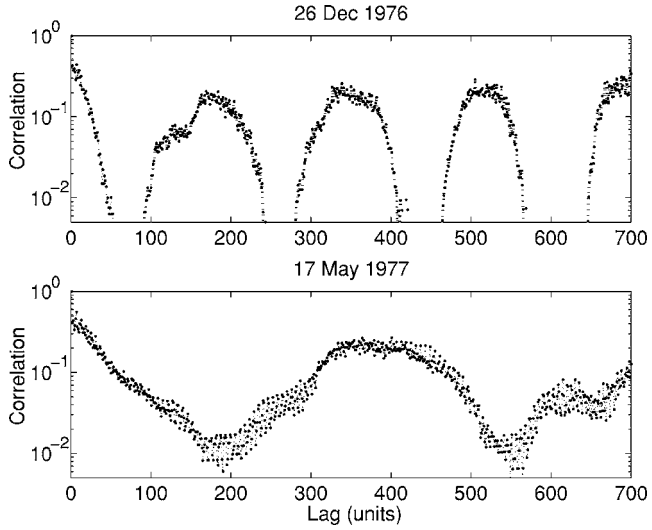

FIG. 6. Global autocorrelation estimates for two songs from the 1976–1977 season. Note that the December song has a period of about 180 units, and the May song of about 400 units. These periods are consistent with the song durations reported in Payne *et al.* (1983). Combining this information with the shorter period shown in Fig. 5, these figures demonstrate that the sequence of units in the songs has multiple periods.

even lower than the values reported in Table II. Thus, the conclusions drawn from Property 3 are still valid. Second, over the window lengths used to obtain $\hat{H}_{\mathrm{SW}}$, the source may be considered locally stationary, as discussed in Sec. I B 7. To justify this assertion, the bottom panel of Fig. 5 enlarges and overlays the two autocorrelation values for $0 \leqslant \lambda \leqslant 40$. This close agreement indicates that the song source may be considered to be locally stationary over windows of 40 units or less. Autocorrelation analysis of other songs exhibited similar agreement over this range of lags. This agreement determined our choice of $\mathcal{I} \in [10,40]$ for the window lengths range in Eq. (26).

The global autocorrelation function $r(x_1^{\ell}, x_1^{\ell}, \lambda)$ shown in Fig. 6 for two songs exhibits a superposition of two oscillations. The shorter period low amplitude oscillations are those of a period of 6–8 units seen in Fig. 5. On the larger timescale used for Fig. 6, these short period oscillations are no longer clearly discernible, but appear as a vertical blur. The larger amplitude oscillation has a period of roughly 180 units for the Dec. 1976 song, and roughly 400 units for the May 1977 song. The other songs analyzed also exhibited a long and short period oscillation. These oscillations demonstrate that the songs possess constraints repeating on segments of 6–8 units in length and also segments hundreds of units in length.

IV. DISCUSSION

Payne *et al.* (1983) previously analyzed a superset of the data used in this study, using human observers of the spectrograms to tabulate the song duration data over 31 day periods. During their Period II of 1976–1977, which includes Dec. 1976, the average song duration was 7.5 min. In their Period VI of the same season, which includes May 1977, the average duration was 13 min. The ratio of their averaged song durations is 1.7. On the other hand, using our correlation analysis, the periods of the Dec. 1976 and May 1977 songs are found to be 180 and 400 units, respectively. The ratio is 2.2. Using an average of 2.5 s per unit, the durations of these songs are approximately 7.5 and 16.7 min. The close agreement between these ratios and durations suggests that the longer period oscillations correspond to their song duration. Additionally, the 6–8 unit oscillations observed in the correlation functions of our Fig. 5 correspond closely to the phrase lengths of 4–10 for most phrases indicated on the spectrograms in Figs. 3 and 4 of Payne *et al.* (1983) and Figs. 6, 8, and 9 of Payne and McVay (1971).

Intriguingly, the entropy and the period of the songs vary over the 1976–1977 singing season. The entropy largely peaks early in the season (Dec. and Jan.) and decreases through the season. Recall from Property 1 of Section I B 2 that a constrained and predictable source has a lower entropy than an unconstrained or unpredictable one. The entropy variations through the season provide an objective quantitative confirmation of several subjective observations made in Payne *et al.* (1983). They observed that "the whales were the most consistent in terms of which themes they included between late 1977 and early 1978… when the number of themes in each song was highest." The relatively low entropy value in May 1977 corroborates this observed consistency in theme selection.

Payne and McVay (1971) defined a transitional phrase to be a phrase occurring between two themes which combines features of both. Some transitional phrases mix units from two adjacent themes in a complex way. Fig. 17 of Payne *et al.* (1983) plots the mean percent of the song duration devoted to transitional phrases. For the 1976–77 season, Period II has the highest proportion of transitional phrases (5%), and the proportion monotonically decreases toward the end of the season. These complicated unusual phrases should increase the total entropy of the song, and the decrease in the proportion of transitional phrases is consistent with the trend found in this analysis of decreasing entropy estimates through the season. Similarly, Payne *et al.* (1983) observed "During the time when the song was least stable in terms of which component themes were present, it contained transitional phrases between all themes; but at the end of the 1976–77 season, when all themes were firmly established, there were no transitional phrases left…. In other words, the song was maximally compartmentalized, organized, and predictable." Again, this qualitative observation is consistent with the general trend of decreasing entropy toward May 1977.

Guinee and Payne (1988) suggested that songs with larger number of themes (longer songs) are often more redundant in phrase structures than shorter songs. Such redundant long songs would be expected to have lower source entropies than shorter songs. Phrased in terms of the entropy analysis, the song lengths determined by the global correlation analysis and the source entropy of the song should be negatively correlated. This would roughly conserve the total amount of information required to remember a song's structure. Our data generally confirm this for the seven independent song sessions analyzed from the 1976–1977 season. (Note that in Table II, there are two segments from the same

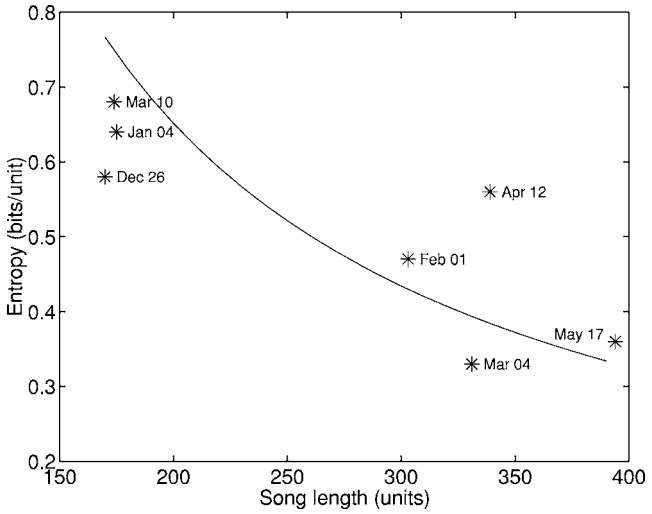

FIG. 7. This scatter plot represents the general trend of increasing song length and decreasing source entropy during the 1976–1977 season observed in seven songs. With the exception of one anomalous singer (Mar. 10), the period of the song increases from Dec. 1976 through May 1977. The source entropy generally decreases through the season, but not as steadily as the period increases. The solid line represents the best fit of a model in which the product of the period and the source entropy is constant for the season. This product, 130.3 bits/song, represents the average total information in the sequence of units in the song.

song session on Mar. 10, and thus these are not really independent observations.) Figure 7 plots the average period versus the source entropy for these seven song sessions. The average period was estimated from the global autocorrelation functions. The product of the song length and source entropy provides a measure of the average amount of total information in the song. One possible relation consistent with Guinee and Payne's (1988) observations is that this product, the bits per song, should be roughly constant through the season. The solid line in Fig. 7 indicates 130.3 bits per song, which is the average of the product of the period and $\hat{H}_{\mathrm{SW}}$ for the song sessions plotted in this figure. Although this relation may be an overly simplistic one, the curve loosely fits the data points, with clear outliers. These observations about the total information are made from a limited number of songs. A more diverse body of data is needed to reach general conclusions.

The results of this study have implications for larger questions comparing the structures of human and nonhuman animal acoustic communications. Some researchers have hypothesized that nonhuman animal acoustic communications lack recursive structures (Chomsky, 1988; Hauser *et al.*, 2002). A recent survey noted that "long-distance hierarchical relations are found in all natural languages, for which, at a minimum, a 'phrase structure grammar' is necessary," (Hauser *et al.*, 2002). The same authors observed that "the core recursive aspect of FLN [the Faculty of Language—Narrow Sense] currently appears to lack any analog in animal communication" and conjectured that "FLN—the computational mechanism of recursion—is recently evolved and unique to our species" (Hauser *et al.*, 2002). This hypothesis is supported by Fitch and Hauser's recent results demonstrating that cotton-top tamarins (*Saguinus oedipus*) are able to parse synthetic stimuli sequences generated by finite-state grammars, but not those generated by a simple recursively hierarchical (phrase structure) grammar (Fitch and Hauser, 2004). Researchers studying human languages describe the property of creating an infinite number of valid signals from a finite set of discrete units as "discrete infinity." Chomsky (1988), in discussing this property, conjectured that "Human language has the extremely unusual, possibly unique, property of discrete infinity" and that "language is based on an entirely different principle than any animal communication system."

The hierarchical structure proposed by Payne and McVay (1971) for humpback whale song challenges these conjectures on the uniquely human nature of long-distance hierarchical relations, and potentially on the uniquely human property of recursion and discrete infinity. Hierarchical grammars may be efficiently represented using recursion, although recursion is not necessarily implied by hierarchy. Additionally, the proposed hierarchical structure and continuous modification of the song raise the possibility that humpback whales can, in theory, create an infinite number of valid songs from the finite set of discrete units. The entropy comparisons in Sec. III A demonstrate that the humpback songs contain structures more complex than first-order Markov models. The lack of stationarity found in the short-term correlation properties of the songs in Sec. III B establishes that the songs cannot be modeled by a higher-order Markov model. The multiple periods on the order 6–8 and hundreds of units found in the autocorrelations are consistent with the hierarchical structure for the songs proposed by Payne and McVay (1971). These multiple periods in the autocorrelation functions illustrate that humpback songs "go beyond purely local structure," demonstrating the sort of "statistical regularities that are separated by an arbitrary number of words" which Hauser *et al.* (2002) asserted are a consequence of natural languages. A hierarchical grammar is a simple and efficient model to produce multiple periods in a song's structure. In light of the previously mentioned inability of some primates to parse hierarchically organized acoustic stimuli (Fitch and Hauser, 2004), this evidence of hierarchical structure in the songs of the evolutionarily more distant humpback whales is intriguing.

It is important to emphasize that we do not claim that humpback songs are a language in the sense recognized by linguists. Hauser *et al.* (2002) define the "conceptual-intentional" component of a language as the property that the different sentences produced as the words are rearranged within the grammatical structure "differ systematically in meaning." There is no evidence that humpback songs satisfy this linguistic requirement. Payne and McVay (1971) made no claims that the large potentially infinite variety of songs produced by the hierarchical ordering of the discrete units have distinct meanings. Tyack (1981) also speculated that each change in the song does not correspond to a change in what the song communicates to other whales. More recently, Payne (1995) explicitly asserted

> "As regards the possible language content of humpback whale songs one can say with certainty only that no one knows whether they contain anything at all that we would equate with language. At the

present state of knowledge, claims to the contrary are simply speculation."

For any one of the components of language enumerated by Hauser *et al.* (2002), it appears possible that there is some animal communication system which possesses that property while lacking one or more of the other linguistic components. Bonobo (*Pan paniscus*) lexigrams appear to satisfy the conceptual-intentional requirement of language (Savage-Rumbaugh *et al.*, 1986), but there are no clear data supporting hierarchical or recursive structure in these animals' communications. The results of the present study are consistent with a hierarchical and recursive structure for humpback song, but as noted above, there is no evidence for humpback song satisfying the conceptual-intentional property of language. This evidence supporting hierarchical structure in humpback song, taken in conjunction with recursion's known efficiency in representing hierarchical structures, raises questions about Hauser *et al.*'s (2002) conjecture that only humans employ recursion to structure their communication signals.

V. CONCLUSION

This paper presents an overview of modern information theory in the context of the analysis of animal communication systems, and then applies these techniques to humpback whale song. The SWML entropy estimator is proposed as a replacement for the i.i.d. and Markov model entropy estimators commonly employed to study animal communications. The SWML estimator is applicable to a much broader class of information sources than the model-based estimators, and its estimates provide valid upper bounds on the entropy for nonstationary sources. The SWML estimator also converges rapidly—an important characteristic for animal communication studies where it is often challenging to obtain long data sequences. In contrast, high-order Markov models risk underestimating the entropy and drawing erroneous conclusions about the source unless extremely large data sets are available.

The entropy and correlation properties are estimated for 16 humpback whale songs. The entropy estimates for both the i.i.d. and first-order Markov models exceed the SWML estimate, indicating that these models fail to capture the structure of the song. The results hold for songs transcribed by two humans and a computer program, controlling for any subjective bias. The SWML estimates indicate that the amount of information carried by the sequence of the units in the song is less than 1 bit per unit. Combining the SWML entropy estimate with the period estimate from the correlation analysis gives an average of roughly 130 bits per song. The correlation analysis demonstrates that songs are nonstationary, but may be considered locally stationary over segment lengths of roughly 40 units. This nonstationarity of the source means that no irreducible empirical Markov model can represent the song structure. The correlation data also indicate that the songs are periodic on two scales of approximately 6–8 and 200–400 units. Such a correlation structure is simply and efficiently produced by hierarchical models—consistent with the Payne and McVay (1971) grammar—but difficult to produce without hierarchy. The entropy and correlation results also provide quantitative confirmation of several observations made by Payne *et al.* (1983) about the song's evolution during the 1976–1977 season. The correlation data demonstrate that the songs possess strong long-distance dependencies of the sort discussed in Hauser *et al.* (2002) as a hallmark of phrase structure grammar. In closing, there is substantial quantitative evidence consistent with the sequence of units in the humpback songs being organized in a hierarchical structure, but equally strong evidence that this sequence is carrying relatively little information.

ACKNOWLEDGMENTS

The authors thank Roger Payne for allowing us to reanalyze the tapes originally made for Payne *et al.* (1983), and Ashley Walker for providing us with unpublished information and programs which inspired us to apply SOMs as classifiers for humpback songs. They also thank Hanspeter Herzel and Tecumseh Fitch for helpful discussions in evaluating and interpreting preliminary results in this study. Hanspeter Herzel also brought Basharin (1959) to our attention. Yoram Bresler and Andrew Solow encouraged us to evaluate the statistical properties of the entropy estimators. Phillip Pendergrass patiently classified hours of humpback song to make this study possible. Two of the authors (J.R.B. and R.S.) gratefully acknowledge the support of the NSF (Ocean Science Career Award No. 9733391). While preparing the final revision of the manuscript, one of the authors (R.S.) was an MIT Rosenblith Fellow and Howard Hughes Medical Institute Predoctoral Fellow. Another author (J.R.B.) received additional support from the Australian-American Fulbright Commission during the revision of the manuscript. ONR Grant No. N00014-00-1-0379 provided the computer facilities used for the entropy estimation. An author (P.L.T.) wishes to acknowledge support from ONR Grant No. N00014-97-1-1031. This is contribution No. 99-1102 from the University of Massachusetts Dartmouth School for Marine Science and Technology and No. 10093 from Woods Hole Oceanographic Institution.

APPENDIX A. PROGRESSION OF INFORMATION ENTROPY AND ITS ESTIMATION

Information entropy was first defined for stationary ergodic Markov sources and the entropy theorem was proved for i.i.d. sources by Shannon (1948). Khinchin (1953, 1957) reinforced Shannon's concepts with rigorous mathematical treatments and proved the entropy theorem for stationary ergodic Markov processes. McMillan (1953) extended the entropy theorem to stationary ergodic sources with convergence in probability. Breiman (1957, 1960) sharpened the convergence of the entropy theorem to convergence with probability one. McMillan (1953) named this theorem the *asymptotic equipartition property*. However, a more intuitive name, the entropy theorem, from Shields (1996), is used in this paper. The coding theorem and its converse for noiseless encoders were first proved by Khinchin (1953, 1957). While Shannon's theory is based entirely on probability theory, Kolmogorov (1965) proposed an alternate measure of com-

plexity which avoids any use of probabilistic concepts. Kolmogorov proposed that length of the shortest binary program that generates the sequence be used as a measure of complexity. Motivated by this approach, the *Lempel-Ziv complexity* (Lempel and Ziv, 1976), *finite state complexity* (Ziv and Lempel, 1978a), and *compressibility* (Ziv and Lempel, 1978b) were defined for individual sequences instead of probabilistic sources. Ziv and Lempel (1978a, 1978b) proved coding theorems and their converses for finite-state complexity and compressibility without using the Shannon entropy, and showed that both concepts are closely related to the source's Shannon entropy by proving Eq. (5) for stationary ergodic sources. They also proposed expanding the concept of entropy to nonstationary sources by using the expected compressibility as an analog of entropy. Verdú and Han (1997) proved that satisfying the entropy theorem is equivalent to satisfying the noiseless coding theorem. They also proposed an extension to the definition of entropy under which some nonstationary sources satisfied the coding theorems. Muramatsu and Kanaya's (1999) extension to the definition of entropy, *almost sure sup entropy*, replaces the limit in Eq. (4) of Definition 2 with a limit supremum. Under this extension of entropy, the coding theorem and its converse hold with probability one for all sources with a consistent probability law.

The basic theory underlying the SWML estimator was developed by Wyner and Ziv (1989). Equation (19) was proven with convergence in probability (Wyner and Ziv, 1989), which was later sharpened to convergence with probability one (Ornstein and Weiss, 1993). Wyner (1993) initially proved the convergence of the average match length [Eq. (20)] for finite memory and fixed window size, which was later expanded in Wyner and Wyner (1995). Kontoyiannis and Suhov (1994) proved a similar result for the Doeblin condition, further developed in Kontoyiannis *et al.* (1998). The SWML entropy estimator was proposed in similar forms by Kontoyiannis, 1997; Kontoyiannis *et al.*, 1998; and Wyner *et al.*, 1998.

[1]This class of encoders are called *noiseless* or *faithful* encoders, meaning that the probability of *any* decoding error is identically zero. The discussion in this paper is limited to this class of encoders, including the well known Ziv and Lempel (1997) encoder. In the literature, faithful encoders are distinguished from the broader class of encoders whose probability of decoding error is finite but can be made arbitrarily small.

[2]A Markov model of order k is equivalent to the $(k+1)$-th order approximation model Shannon (1948, 1951) used for his experiments. In particular, the zeroth order Markov model is equivalent to the i.i.d. model or Shannon's first-order model.

[3]Doeblin condition: There exists an integer $k \geq 1$ and a real number $\gamma \in (0,1)$ such that $\Pr\{X_i = x_i | X_{-\infty}^{i-k}\} \geq \gamma$ for all $x_i \in \mathcal{A}$ at all time i with probability one.

[4]The number of parameters of a Markov model and the required sequence length grow with the number of possible transitions, not combinations. An erroneous guideline using the number of combinations appears in McCowan *et al.* (1999). They claimed that their kth-order Markov models had $C(|\mathcal{A}|, k+1) = |\mathcal{A}|!/(k+1)!(|\mathcal{A}|-(k+1))!$ parameters and required the sequence length to exceed this number. The correct number of parameters is actually $|\mathcal{A}|^{(k+1)}$, and thus much longer sequences are required. For example, when $|\mathcal{A}|=20$, a second-order Markov model has 8000 parameters, almost an order of magnitude more than the 1140 predicted by C(20,3). Entropy estimates from incorrect models or insufficient data are uninterpretable.

[5]The bias of an estimator is defined as $\mathbf{E}\{\hat{\theta}\} - \theta$. A negative bias indicates that the expected estimate is lower than the true value.

[6]In contrast to Payne and McVay's (1971) observation, we found some units significantly longer than 6 s. Therefore, we allocated 12.8 s for the time axis of the spectrogram.

Basharin, G. P. (**1959**). "On a statistical estimate for the entropy of a sequence of independent random variables," Theor. Probab. Appl. **4**, 333–336.

Beecher, M. D. (**1989**). "Signalling systems for individual recognition: an information theory approach," Anim. Behav. **38**, 248–261.

Breiman, L. (**1957**). "The individual ergodic theorem of information theory," Ann. Math. Stat. **28**, 809–811 also **31**, 809–810 (**1960**).

Breiman, L. (**1960**). "A correction to 'The individual ergodic theorem of information theory' ", Ann. Math. Stat. **31**, 809–810.

Chomsky, N. (**1956**). "Three models for the description of language," IRE Trans. Inf. Theory **2**, 113–124.

Chomsky, N. (**1988**). *Language and Problems of Knowledge: The Managua Lectures* (MIT Press, Cambridge, MA).

Cover, T. M., and King, R. C. (**1978**). "A convergent gambling estimate of the entropy of English," IEEE Trans. Inf. Theory **24**, 413–420.

Cover, T., and Thomas, J. (**1991**). *Elements of Information Theory* (Wiley, NY).

D'Vincent, C. G., Nilson, R. M., and Hanna, R. E. (**1985**). "Vocalization and coordinated feeding behavior of the humpback whale in southeastern Alaska," Sci. Rep. Whales Res. Inst. **36**, 41–47.

Duda, R. O., Hart, P. E., and Stork, D. G. (**2001**). *Pattern Classification*, 2nd ed. (Wiley, NY).

Fitch, W. T., and Hauser, M. D. (**2004**). "Computational constraints on syntactic processing in a nonhuman primate," Science **303**, 376–380.

Gentner, T. Q., and Hulse, S. H. (**1998**). "Perceptual mechanisms for individual vocal recognition in European starlings, *Sturnus vulgaris*," Anim. Behav. **56**, 579–594.

Good, I. J. (**1953**). "The population frequencies of species and estimation of population parameters," Biometrika **40**, 237–264.

Guinee, L. N., and Payne, K. B. (**1988**). "Rhyme-like repetitions in songs of humpback whales," Ethology **79**, 295–306.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (**2002**). "The faculty of language: What is it, who has it, and how did it evolve?," Science **298**, 1569–1579.

Haykin, S. (**1999**). *Neural Networks: A Comprehensive Foundation*, 2nd ed., (Prentice–Hall, Englewood Cliffs, NJ).

Janik, V. M. (**1999**). "Pitfalls in the categorization of behaviour: A comparison of dolphin whistle classification methods," Anim. Behav. **57**, 133–143.

Janik, V. M., and Slater, P. J. B. (**1997**). "Vocal learning in mammals," in *Advances in the Study of Behavior, Volume 26*, edited by P. J. B. Slater, C. Snowdon, J. Rosenblatt, and M. Milinksi (Academic, NY), pp. 59–99.

Jankowski, C. R., Vo, H.-D. H., and Lippmann, R. P. (**1995**). "A comparison of signal processing front ends for automatic word recognition," IEEE Trans. Speech Audio Process. **3**, 286–293.

Ketten, D. R. (**1997**). "Structure and function in whale ears," Bioacoustics **8**, 103–135.

Khinchin, A. I. (**1953**). "The entropy concept in probability theory," Usp. Mat. Nauk **8**, 3–20 [English translation in Khinchin (1957)].

Khinchin, A. I. (**1957**). *Mathematical Foundations of Information Theory* (Dover, NY).

Kohonen, T. (**2001**). *Self-Organizing Maps*, 3rd ed., (Springer, NY).

Kolmogorov, A. N., (**1965**). "Three approaches to the quantitative definition of information," Probl. Inf. Transm. **1**, 1–7.

Kontoyiannis, I. (**1997**). "The complexity and entropy of literary styles," NSF Technical Report No. 97, Department of Statistics, Stanford University, Palo Alto, CA, June 1996/October 1997.

Kontoyiannis, I., and Suhov, Y. M. (**1994**). "Prefixes and the entropy rate for long-range sources," in *Probability, Statistics and Optimization*, edited by F. P. Kelly (Wiley, NY), pp. 89–98.

Kontoyiannis, I., Algoet, P. H., Suhov, Y. M., and Wyner, A. J. (**1998**). "Nonparametric entropy estimation for stationary processes and random fields, with applications to English text," IEEE Trans. Inf. Theory **44**, 1319–27.

Lempel, A., and Ziv, J. (**1976**). "On the complexity of finite sequences," IEEE Trans. Inf. Theory **22**, 75–81.

Levitin, L. B., and Reingold, Z. (**1994**). "Entropy of natural languages: Theory and experiment," Chaos, Solitons Fractals **4**, 709–743.

Linde, Y., Buzo, A., and Gray, R. M. (**1980**). "An algorithm for vector quantizer design," IEEE Trans. Commun. **28**, 84–95.
MacKay, D. M. (**1972**). "Formal analysis of communicative processes," in *Nonverbal Communication*, edited by R. A. Hinde (Cambridge University Press, Cambridge).
Marton, K., and Shields, P. (**1994**). "Entropy and the consistent estimation of joint distributions," Ann. Prob. 22, 960–977.
Marton, K., and Shields, P. (**1996**). "Entropy and the consistent estimation of joint distributions," Ann. Prob. 24, 541–545.
McCowan, B., Hanser, S. F., and Doyle, L. R. (**1999**). "Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires," Anim. Behav. **57**, 409–419.
McMillan, B. (**1953**). "The basic theorems of information theory," Ann. Math. Stat. **24**, 196–219.
Miller, G. A. (**1954**). "Communication," Annu. Rev. Psychol. **5**, 401–420.
Miller, G. A., and Chomsky, N. (**1963**). "Finitary models of language users," in *Handbook of Mathematical Psychology*, edited by R. D. Luce, R. R. Bush, and E. Galanter (Wiley, NY), pp. 419–492.
Miller, P. J. O., Biassoni, N., Samuels, A., and Tyack, P. L. (**2000**). "Whale songs lengthen in response to sonar," Nature (London) **405**, 903.
Muramatsu, J., and Kanaya, F. (**1999**). "Almost-sure variable-length source coding theorems for general sources," IEEE Trans. Inf. Theory **45**, 337–342.
Noad, M., Cato, D. H., and Bryden, M. M. (**2000**) "Cultural revolution in whale songs," Nature (London) **408**, 537.
Ornstein, D. S., and Weiss, B. (**1993**). "Entropy and data compression schemes," IEEE Trans. Inf. Theory **39**, 78–83.
Patil, G. P., and Taillie, C. (**1982**). "Diversity as a concept and its measurement," J. Am. Stat. Assoc. **77**, 548–567 (with discussions).
Payne, R. (**1995**). *Among Whales* (Scribner's Sons, NY).
Payne, R. S., and McVay, S. (**1971**). "Songs of humpback whales," Science **173**, 587–597.
Payne, K., Tyack, P., and Payne, R. (**1983**). "Progressive changes in the songs of humpback whales (*Megaptera novaeangliae*): A detailed analysis of two seasons in Hawaii," in *Communication and Behavior of Whales*, edited by R. Payne, AAAS Selected Symposium 76 (Westview, Boulder, CO), pp. 9–58.
Peet, R. K. (**1974**). "The measurement of species diversity," Annu. Rev. Ecol. Syst. **5**, 285–307.
Pitton, J. W., Wang, K., and Juang, B.-H. (**1996**) "Time-frequency analysis and auditory modeling for automatic recognition of speech," Proc. IEEE **84**, 1199–1215.
Savage-Rumbaugh, S., McDonald, K., Sevcik, R. A., Hopkins, W. D., and Rubert, E. (**1986**), "Spontaneous symbol acquisition and communicative use by pygmy chimpanzees (*Pan paniscus*)," J. Exp. Psychol. **115**, 211–235.
Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423.
Shannon, C. E. (**1951**). "Prediction and entropy of printed English," Bell Syst. Tech. J. **30**, 50–64.
Shields, P. C. (**1996**). *Ergodic Theory of Discrete Sample Paths* (American Mathematical Society, Providence, RI).
Slater, P. J. B. (**1973**). "Describing sequences of behavior," in *Perspectives in Ethology*, edited by P. P. G. Bateson, and P. H. Klopfer, (Plenum Press, New York), Vol. 1, pp. 131–153.
Tyack, P. L. (**1981**). "Interactions between singing Hawaiian humpback whales and conspecifics nearby," Behav. Ecol. Sociobiol. **8**, 105–116.
Tyack, P. L. (**1998**). "Acoustic communication under the sea," in *Animal Acoustic Communication: Sound Analysis and Research Methods*, edited by S. L. Hopp, M. J. Owren, and C. S. Evans (Springer, Berlin), pp. 163–220.
Tyack, P. L., and Sayigh, L. S. (**1997**). "Vocal learning in cetaceans," in *Social Influences on Vocal Development*, edited by C. Snowdon, and M. Hausberger (Cambridge University Press, Cambridge), pp. 208–233.
Verdú, S., and Han, T. S. (**1997**). "The role of the asymptotic equipartition property in noiseless source coding," IEEE Trans. Inf. Theory **43**, 847–857.
Walker, A., Fisher, R. B., and Mitsakakis, N. (1996). "Singing maps: Classification of whalesong units using a self-organizing feature mapping algorithm," Research Paper No. 833, Department of AI, University of Edinburgh, UK.
Winn, H. E., and Winn, L. K. (**1978**). "The song of the humpback whale *Megaptera novaeangliae* in the West Indies," Mar. Biol. (Berlin) **47**, 97–114.
Wyner, A. J. (1993). "String matching theorems and applications to data compression and statistics," Ph.D. dissertation, Department of Statistics, Stanford University, Palo Alto, CA.
Wyner, A. D., and Ziv, J. (**1989**). "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," IEEE Trans. Inf. Theory **35**, 1250–1258.
Wyner, A. D., and Wyner, A. J. (**1994**). "The sliding window Lempel–Ziv algorithm is asymptotically optimal," Proc. IEEE **82**, 872–877.
Wyner, A. D., and Wyner, A. J. (**1995**). "Improved redundancy of a version of the Lempel–Ziv algorithm," IEEE Trans. Inf. Theory **41**, 723–731.
Wyner, A. D., Ziv, J., and Wyner, A. J. (**1998**). "On the role of pattern matching in information theory," IEEE Trans. Inf. Theory **44**, 2045–2056.
Ziv, J., and Lempel, A. (**1977**). "A universal algorithm for sequential data compression," IEEE Trans. Inf. Theory **23**, 337–343.
Ziv, J., and Lempel, A. (**1978a**). "Coding theorems for individual sequences," IEEE Trans. Inf. Theory **24**, 405–412.
Ziv, J., and Lempel, A. (**1978b**). "Compression of individual sequences via variable-rate coding," IEEE Trans. Inf. Theory **24**, 530–536.