Methods for Automatically Extracting Bioacoustic Structure

Peter Rickwood
University of Technology, Sydney, Australia

Andrew Taylor*
School of Computer Science and Engineering
University of New South Wales, Australia
(Dated: March 26, 2007)

This paper presents mathematical methods for automatically extracting and analysing bioacoustic signals. Automatic techniques are described for isolation of target signals from background noise, extraction of features from target signals and unsupervised classification (clustering) of the target signals based on these features. The only user-provided inputs, other than raw sound, are 6 control parameters. In particular, the number of signal categories is determined automatically. The techniques, applied to noisy hydrophone recordings of Humpback Whales (*Megaptera novaeangliae*), achieve good results, suggesting they are sufficiently general to be applicable in many other bioacoustic settings.

I. INTRODUCTION

Almost all analysis of bioacoustic signals is done with substantial human guidance. This can be extremely time consuming. It can also raise serious concerns regarding subjectivity. Lack of quantification can make replication or comparison of manual analyses difficult. There are two principal tasks in extracting the structure of bioacoustic signals:

1. **Signal segmentation:** The separation of the signal component potentially originating from one or more target sources from the remainder of the signal (the background). The target sources may be a single individual, multiple conspecific individuals or multiple individuals from multiple species. The signal background may include: bioacoustic signals from non-target individuals of the same species; bioacoustic signals from other species and non-bioacoustic sounds.

2. **Signal characterization:** The characterization of the signal component from the target individual(s) isolated in the previous stage. This characterization will involve reduction of the signal into a more useful and typically much more compact form.

This paper describes automatic techniques for each task and present the successful results of applied these methods to characterize a noisy set of complex bioacoustic signals - Humpback Whale *Megaptera novaeangliae* vocalizations. An open-source software implementation of the techniques described in this paper is available from the authors.

*Electronic address: andrewt@cse.unsw.edu.au

II. TARGET SIGNAL SEGMENTATION

Signal segmentation is a problem common to many domains. For example, it has been well studied for human speech recognition ([8, 36]). Domain-specific characteristics presumably explain why these techniques have not been transferred to bioacoustic applications. Some researchers have isolated bioacoustic signals manually ([18, 30]), or avoided the issue altogether ([15]).

It is desirable to minimize assumptions on the amplitude, duration, or frequency characteristics of the vocalizations as these can vary greatly between target species.

Our approach is energy-based. It assumes intervals of the signal which contain target components will contain more energy within a certain frequency range than other intervals. Further assumptions might produce better detection of target components but risk the reduction of generality.

A. Signal Segmentation, Step 1: Vector Extraction

The first step in this extraction process is to partition the recording into fixed-length frames. Background noise is assumed to vary sufficiently slowly that it is almost constant within frames. For estimation of background noise, it is assumed that target vocalizations occupy at most 70% of each frame. The technique thus depends on target signals being interspersed by sufficient periods of background noise, which is typical of bioacoustic environments. The exact choice of frame length is relatively unimportant if it significantly exceeds the duration of vocalizations. Frame length is a manually chosen parameter - we have not explored automatic determination of frame length but this may be possible. The users must also choose parameters for FFT size, FFT step and a frequency range appropriate for the target signals.

For humpbacks we use a frame length of 20 seconds, a Hann-windowed 1024 point FFT with a 1/3 window

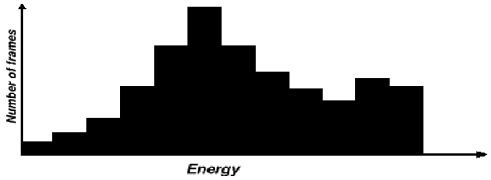

FIG. 1: The combined energy histogram of target and noise signal is usually a combination of a low energy, approximately symmetric noise distribution, and a complex distribution covering signal energy.

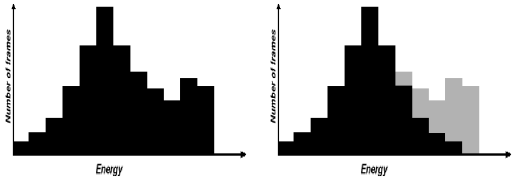

FIG. 2: An illustration of how it is possible, given an energy histogram, to separate signal and noise distributions by assuming that the distribution covering noise is symmetric. For each possible energy threshold, the black area to the right of that threshold is proportional to the expected number of incorrectly labelled noise samples, and the grey area to the left of the threshold is proportional to the number of incorrectly labelled target samples.

step and a 90-5600 Hz frequency range. The majority of humpback signal energy is in the 0-6 kHz range ([11, 12, 17, 28, 29, 42, 44]). We exclude the 0-90 Hz range because ambient noise typically dominates this range. Our humpback data set has been digitized at 32 kHz. A 20 second frame with 1024 point FFT and 341 sample FFT step produces 1876 values for each frame.

B. Signal Segmentation, Step 2: Differentiating signal and noise

To begin with, a rough grouping is made based purely on band-limited energy. That is, for intervals where energy is greater than some threshold E, the target signal is considered to be present, otherwise it is considered absent. Calculation of the threshold value E is done through some simple histogram analysis. All values in the frame are used to create a histogram which describes the distribution of band limited energy. Figure 1 shows a typical histogram.

Making some assumptions about the general form of this energy histogram will allow us to calculate a threshold energy E. As a start, it is assumed that noise samples are distributed according to a distribution f, and signal samples are distributed according to g. The energy histogram of the band-limited energy samples in each frame can then be described by the distribution $(f + g)$. The following further assumptions are made:

1. Samples drawn from g will in general have greater energy than those drawn from f. i.e. $E(g(X)) > E(f(X))$.
2. f is approximately symmetric.
3. The probability of an observation drawn from g having less energy than the mode of f is 0. (i.e. $E(g(X) < \text{mode}(f) = 0)$
4. The mode of $(f + g)$ is the same as the mode of f.

The procedure is then as follows:

1. Estimate the mode of $(f + g)$, which also gives us the mode of f (from assumption 4).
2. Reconstruct f from its known mode. This can be done using assumptions 1,2 and 3.
3. For all possible thresholds, estimate the number of misclassified samples if all samples below that threshold are deemed noise and all above deemed signal. Since we know $(f + g)$ and f, we also know g, which means this is a trivial calculation. See Figure 2.

Once a threshold energy is calculated, each sample in the time series can be tentatively labelled as either target (if it contains more energy than the threshold), or background (if less). Detection is improved by making a further assumption:

- There is some minimum time period, of length d seconds, which must separate an interval of ambient noise from an interval of a target signal (i.e. a vocalization). That is, if a signal/noise interval is detected, no noise/signal interval can occur before d seconds has elapsed.

The assumption allows us to apply a temporal smoothing procedure, where each sample is classified as target only if the majority of other samples in a d second neighbourhood around it are also classified as target. This smoothing is applied repeatedly until no changes take place. The separation period can be small relative to the expected signal length – for humpbacks, we use $d = 0.05$ – despite the fact that most humpback vocalizations are much longer than this.

The histogram analysis, followed by temporal smoothing, is performed for each frame in the recording, so that there is a complete classification of each interval in the recording into either target or noise. Each contiguous time interval labelled as target is regarded as a single distinct vocalization.

Figure 3 shows this technique as applied to a frame containing a series of vocalizations. All 15 vocalizations are successfully detected, with only 1 false positive (the

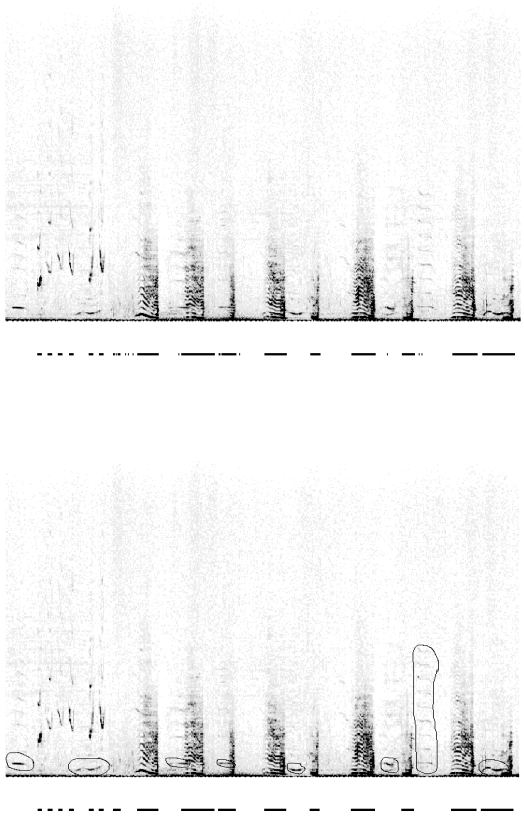

FIG. 3: The black line under each spectrogram indicates which intervals have been classified as target. The left figure is based on energy alone, and the right figure shows the effect of smoothing on this classification. Notice that good results are obtained on this 20 second signal despite the presence of a second whale singing (circled on right).

seventh interval from the left), and some artificial elongation of the ninth and sixteenth intervals. The technique successfully excludes most ambient noise and, more impressively, the fainter vocalizations of a second (more distant) whale, which are circled in the bottom figure. This is sufficiently accurate to allow successful unsupervised classification.

III. UNSUPERVISED CLASSIFICATION OF SOUNDS

There has been a variety of work on (supervised) classification of bioacoustic signals into (predetermined) categories, e.g. ([15, 30, 41]). This paper does not address supervised classification. It presents techniques for unsupervised classification (clustering) – classification without predetermined categories.

Applications of unsupervised classification to bioacoustic signals has been limited, e.g.([6, 18, 25]), and typically with a highly target-specific approach. Unsupervised classification can be viewed as constructing a model for a set of signals. For some purposes measurement of several signal parameters may sufficiently model the signal (see [2, 15, 17, 20]). Analysis of more complex bioacoustic signals requires more complex models. This paper details general methods to automatically build a class of such models. These methods handle robustly errors in signal segmentation, allowing automatic signal segmentation to be employed and hence fully automatic signal analysis.

Unsupervised classification techniques for sound are not well established, because the temporal nature of a sound is difficult to capture with standard attribute-value clustering techniques like *AutoClass* ([7]) or *SNOB* ([43]). This has resulted in the use of a large number of more exotic techniques being used, such as Kohonen networks ([21]), Adaptive Resonance Theory Networks ([4]), various clustering techniques (see [6, 18]), and many others. Often, the user must specify suitable value for a number of parameters to obtain good performance from these classifiers. The technique that we describe avoids the need for extensive parameter-tweaking by using a penalty function (inspired by information theory) which finds a trade-off between model complexity and model fit to data.

A. Feature extraction

Working with raw sound data is difficult, so after target signals have been isolated, the next step is to extract from the raw sound data some higher level features that accurately describe the time varying frequency and amplitude characteristics of the signal. We use simple feature vectors extracted from the same power spectrum vectors calculated during signal segmentation. Each FFT produces a power spectrum vector $\mathbf{p} = [p_1 \dots p_n]$. From this vector, a simple feature vector $\mathbf{q} = [q_1 \dots q_m]$ is computed, where each q_i is the sum of one or more of the elements of $\mathbf{p}$.

The parameters, n, m and the computation of q_i are user specified. For humpbacks, we use $n = 512$ and $m = 36$ with q_i corresponding to exponentially increasing frequency band sizes, with a small exponent. Specifically, the number of FFT elements in each feature vector element is given by the series $[\lfloor 2^{1.05} \rfloor, \lfloor 2^{1.05^2} \rfloor, \lfloor 2^{1.05^3} \rfloor, \dots]$. For example, $q_5 = p_{12} + p_{13}$ (375-406hz) and $q_{33} = p_{140} + p_{141} + \dots + p_{148}$ (4375-4625hz)

This yields for every target signal a sequence of feature vectors describing the time-varying power spectra of the recording over that interval. These are the basis for all further analysis. More sophisticated feature vectors are commonly used in speech recognition (see [13, 14]), and similar feature vectors might improve performance with humpbacks, for example, but at the cost of simplicity and perhaps generality.

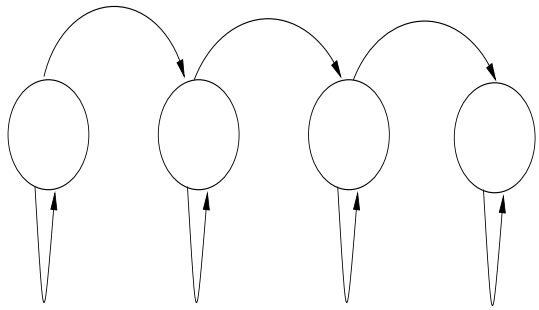

FIG. 4: A Bakis Hidden Markov Model.

B. Vector Quantization

Each vocalization O (detected by the technique in Section II), consists of a sequence of feature vectors (i.e. $O = [\mathbf{q_1} \dots \mathbf{q_n}]$, with each $\mathbf{q_i}$ being a feature vector $\mathbf{q_i} = [q_{i_1} \dots q_{i_m}]$), extracted as described above. Each such sequence of vectors is reduced to a sequence of integers through the process of *vector quantization* ([3, 24, 35]). Vector quantization essentially clustering vectors in a multi-dimensional space, and replaces each vector with the index of the cluster centroid that the vector is closest to, under some distance metric (usually squared Euclidean distance). This process allows us to describe each vocalization as a sequence of integers $[i_1 \dots i_n]$ rather than a sequence of feature vectors. The vector quantization/clustering required to do this is performed on all feature vectors extracted (i.e. on every feature vector of every target sound detected in the target segmentation step).

The codebook size (number of cluster centroids) is specified by the user, for humpbacks we used 512. It is likely it also could be well determined by an adaptive/automatic approach.

C. Hidden Markov Modelling of Sound

Hidden Markov Models (HMMs) [19, 33] are better known as a tool used in supervised classification of sound, especially speech recognition ([23, 32, 34, 35]), but the ability of HMMs to model the time-varying nature of sound signals also makes them suitable for unsupervised classification.

Traditionally, Bakis HMMs (HMMs with transitions only allowed from left to right, see Figure 4) have been used to model time-varying sounds.

Vector quantization has reduced each vocalization to a sequence of integers. The integers become the alphabet of our HMM. Given a set of integer sequences $\{O_1 \dots O_r\}$, where each O_i is some integer sequence of symbols $[i_1 \dots i_n]$, each of which describes a single vocalization through time, we can calculate the following:

1. $P(\{O_1 \dots O_r\}|M)$: the conditional probability of observing the sequences, given a particular Hidden Markov Model M.
2. Calculate M_{opt}, the HMM that locally maximizes $P(\{O_1 \dots O_r\}|M)$ (using Baum-Welch re-estimation).

For now, assume there are k different types of vocalization in the data set – a method for determining k will be presented later. k HMMs ($\{M_1 \dots M_k\}$) modelling these types of vocalization can be constructed using this algorithm:

1. Partition all observed sequences into k disjoint sets. Call these $\{C_1 \dots C_k\}$.
2. For each $C_i \in \{C_1 \dots C_k\}$, calculate M_{opt}, the HMM that locally maximizes $P(C_i|M)$. This HMM is said to describe the elements of C_i. In this way, each HMM is associated with a disjoint subset of sequences.
3. For every vocalization (i.e. each O_i), calculate $P(O_i|M)$ for each HMM $M \in \{M_1 \dots M_k\}$. O_i is then assigned to the subset C_i associated with the HMM that maximizes $P(O_i|M)$.
4. If any sequences changed their subset membership, goto step 2.

In practice, waiting for complete termination (i.e. no change in subset membership) is inefficient, and, indeed, convergence is not guaranteed, so the iterations cease when only a small number ($\approx 1\%$) of sequences have an altered subset membership.

Given any particular vocalization O_i, it is possible to determine its type by calculating which of these HMMs maximizes $P(O_i|M)$. This in itself is a significant development, as it is now possible to use the power of HMMs to model time varying bioacoustic signals.

The above algorithm crucially relies on the value of k. The following section describes how this is determined using an information-theory derived penalty function.

1. MDL based classification

Minimum message length (MDL) encoding ([1, 16, 22, 26, 27, 37, 38, 43]) is an information theoretic approach to balancing a model's fitting of observed data against its complexity. In essence MDL encoding states that, for a set of models $\{m_1, m_2, \dots, m_n\}$, each of which is an alternative way of describing a probability distribution over observed data, one can assess the quality of any particular model m_i by measuring the length of the binary string required to describe the model and the observed data. The best model is the one that minimises the length of this description string. MDL encoding has been successfully employed in widely-used supervised ([31]) and unsupervised ([43]) classification systems.

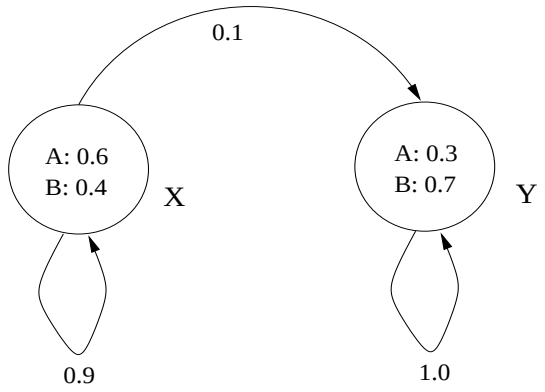

FIG. 5: A simple 2-state HMM.

We wish to choose a value for k, the number of vocalization types, that minimises the number of bits, usually termed the *transmission cost*, required to encode the vocalizations (the observed data), and the models that describe those vocalizations. This requires an encoding scheme for HMMs.

Suppose we have a simple 2-state HMM, M, shown in Figure 5, with observable symbols $\{A, B\}$ and a single observed sequence: $[A, A, A, B, B, A, B]$.

The transmission cost consists of two parts: The model cost $C(M)$, which is the number of bits to encode the model; plus the message cost $C([A, A, A, B, B, A, B]|M)$, which is the number of bits to encode the observed symbols $[A, A, A, B, B, A, B]$, using the model M.

Communicating the model consists of communicating the transition and emission probabilities for each state of the model. The optimal encoding for an event with probability p is $-\log_2 p$ bits [40]. For the first state of the model shown in Figure 5, these are:

Symbol	Probability	Codeword length
A	0.6	0.74 bits
B	0.4	1.32 bits

Transition	Probability	Codeword length
X → X	0.9	0.15 bits
X → Y	0.1	3.32 bits

The cost of communicating the transition probabilities for this state is 3.47 bits, and the cost of communicating the emission probabilities is 2.06 bits. So the total transition cost for that state is 5.53 bits. This process is repeated for each state to get the total cost of specifying the entire model. For this example, it's 7.78 bits, since state 2 takes 2.25 bits.

It is straightforward to calculate the probability $P(O|M)$ of observing some sequence O for a particular HMM M ([33]) allowing us to us to calculate the message cost $-\log_2 P(O|M)$, the number of bits required to communicate the observed data given the model.

IV. DETERMINING THE NUMBER OF CLUSTERS

The ability to calculate the transmission cost of a HMM and the data it describes allows us to compare alternative ways of classifying observations. For example, one could initially set k to 1 (i.e. have a single category of vocalization, modeled by a single HMM M_1^1), and calculate $C(M_1^1)$ and then calculate $\sum_{i=1}^{r} C(O_i|M_1^1)$ for all observations $O_1 \ldots O_r$. The sum $C(M_1^1) + \sum_{i=1}^{r} C(O_i|M_1^1)$ is the total transmission cost for model and data. We could then repeat for increasing values of k, and see how the total transmission cost varies. Choosing the value of k that minimises total transmission cost would give an estimate of k. While this approach is feasible, it can be computationally expensive, and so we now describe a more efficient method of achieving the same end – that is, estimating a value of k that has good fit to data but is sufficiently small so that the categories are intelligible to a human viewing the results. Note that, in a purely statistical sense, simpler models are not always preferable ([39],[10]), but they do have the (non-statistical) advantage of being easier to interpret.

1. Initially, a 'guess' is made for the initial number of classes, k_{guess}. Vocalizations are assigned to classes at random. A HMM is then generated for each class, which is then trained (using Baum-Welch re-estimation) on the members of the class. This will give k_{guess} HMMs, each of which models some disjoint subset of the instances that are being classified.

2. Iterative expectation-maximization ([9]) is now performed until class membership is stable:

   (a) M: Train each HMM using the members that belong to its class. This is done with Baum-Welch re-estimation.

   (b) E: Re-estimate class membership for each vocalization, based on the newly trained HMMs by assigning each vocalization to the class whose HMM maximizes $P(O|M)$.

   (c) goto (1) if any reassignments took place in (2).

3. For each of the k_{guess} HMMs (call them $\{M_1 \ldots M_{k_{\text{guess}}}\}$), calculate whether removing that HMM results in a reduction in total transmission cost.

4. Test to see if splitting any classes results in an information gain (i.e. a reduced number of bits). That is, for every HMM M, modeling a set of observations O, partition O into two random subsets (O', O''), generate a HMM for each subset (M', M''), and perform HMM training and re-estimation (as in 2), until subset membership is stable. Once this is done the transmission cost of

the original grouping (i.e. $C(M)+C(O|M)$) is compared with that of the proposed split ($C(M') + C(O'|M') + C(M'') + C(O''|M'')$). Whichever has the lower transmission cost is retained. That is, if $C(M') + C(O'|M') + C(M'') + C(O''|M'') < C(M) + C(O|M)$, M is replaced with M' and M'', otherwise M is retained and M' and M'' are discarded. If any of these 'splits' do occur, the splitting test is applied recursively to each new subset, until there is no information gain in splitting. After this process is complete, let the remaining number of classes be k''. Figure 6 demonstrates the recursive splitting of a single class (covering four observations $o2, o4, o7, o9$) into three separate classes.

5. Consider there now to be k'' remaining classes/HMMs, with each still representing a disjoint subset of instances. At this point, the algorithm repeatedly performs the expectation/maximization described in step 2, until no changes to class membership takes place. This second re-estimation step helps to move out of undesirable locally optimal solutions produced by the greedy split/join procedure.

6. Perform one final time the procedure described in step 3 – removing HMMs while there is an information gain in doing so. Let the remaining number of clusters be k'''.

7. This completes the classifying procedure, with k''' being the number of classes 'decided' on by the algorithm.

One objection that may be raised is that the technique for automatic grouping of vocalizations requires an initial 'guess', by the user, of the number of distinct types of call. However, our initial experiments with humpback vocalizations seem to indicate that this initial guess in unimportant, playing little role in the final number of classes determined by the technique. In other words, the technique is not sensitive to this initial guess. Figure 7 illustrates.

In addition, we have thus far avoided specifying how we determine the number of states in the HMM that models each call category. While this could be determined through some automatic scheme, we choose the simpler and faster method where the number of states is proportional to the mean length of the vocalizations modeled by the HMM. As the number of states is proportional to call length, the user must specify the constant parameter.

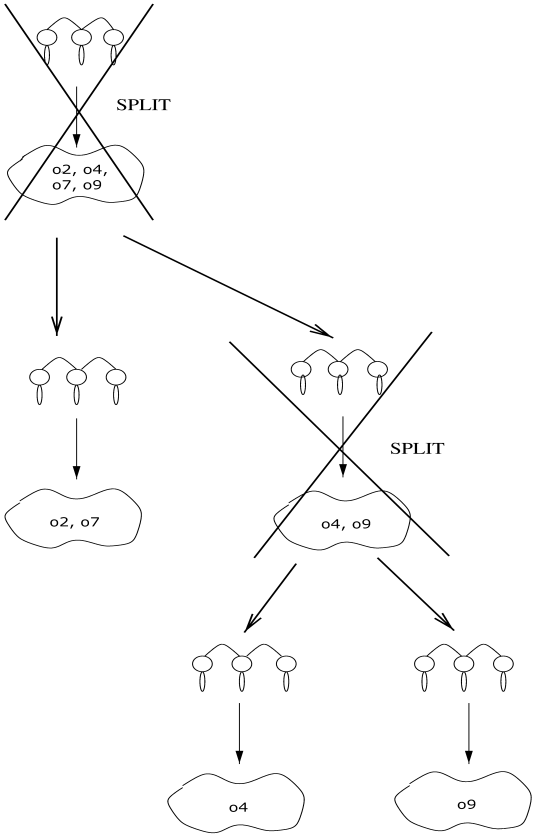

FIG. 6: Recursive splitting of a single class.

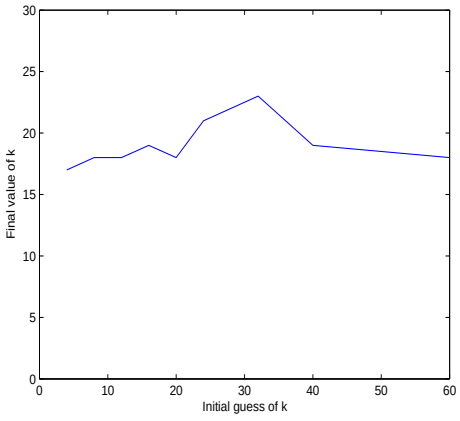

FIG. 7: Final number of clusters against initial number of seed clusters.

V. RESULTS

We applied the techniques (signal segmentation, feature extraction, clustering) to 11 hours of humpback recordings. The recordings consisted of 9 separate recordings, each of between 54 and 90 minutes, taken on separate days via hydrophone buoy off the coast of North Stradbroke Island in Queensland, Australia – a known migration path of humpbacks. Each recording captures the vocalizations of an individual humpback as they pass near the buoy. The depth and orientation of the whale to the buoy vary through time. Wave and wind noise also vary throughout the recordings. Boat motors are recorded, as are other (more distant) whales. Noise levels vary by over 10 dB through the 11 hours of recordings,

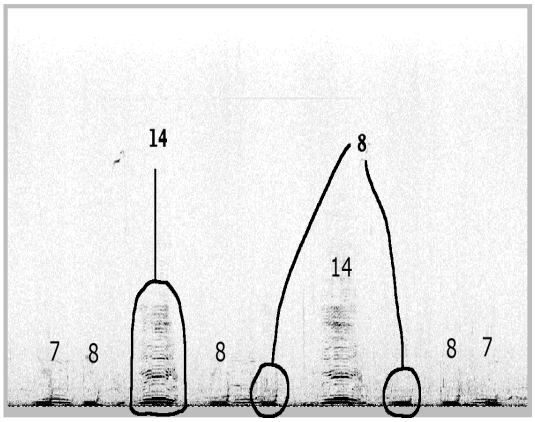

FIG. 8:

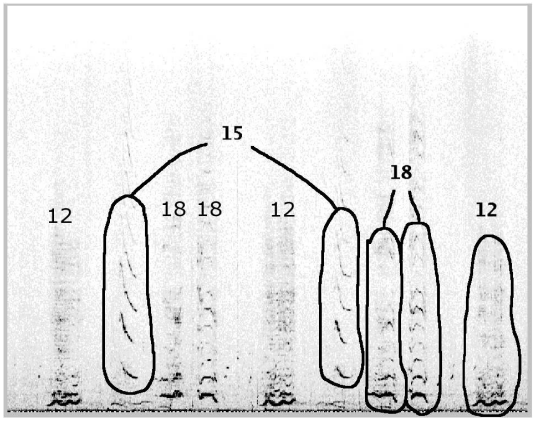

FIG. 9:

but in any 20 second interval, variation in background noise is generally less than 3 dB. We performed no filtering or other pre-processing to minimise the effect of noise sources. At no stage did we intervene in the process of signal segmentation, extraction, or classification. Thus, we did not vet the results of the segmentation algorithm before running the feature extraction and clustering. The steps thus form a single processing pipeline. Some indication of the accuracy of the segmentation technique is given by Figure 3. More convincing evidence of the ability of the segmentation technique to do a reasonable job is provided by the results of the signal classification algorithm – since the classification/clustering procedure is built on top of the segmentation and extraction procedures, it cannot perform well if the segmentation procedure does not.

The unsupervised classification technique described produced 19 distinct call types after processing $\approx$ 11 hours of humpback recordings. As noted, the segmentation technique described in Section II is not perfect, and a cursory inspection of the 19 call types identifies 9 of them as describing 'noise' clusters, rather than actual humpback vocalizations – leaving 10 distinct types of humpback call identified. Despite the fact that the signal segmentation procedure produces a number of false positives, the ability of the algorithm to group mistakenly isolated noise into distinct categories based on their spectral characteristics is pleasing, and is an indication of the robustness of the technique. Figures 8 to 15 show randomly chosen extracts from the 11 hours of recording with labels attached to indicate the 'category' to which each vocalization has been assigned. To keep the figures clear, 'noise' categories mentioned above are excluded. A few representative samples of each signal category are circled, with the rest simply having their category listed above the call. Some vocalizations that a human might consider similar are grouped separately (for example, 10 and 9), but we can see that the technique overall does a good job of grouping similar vocalizations.

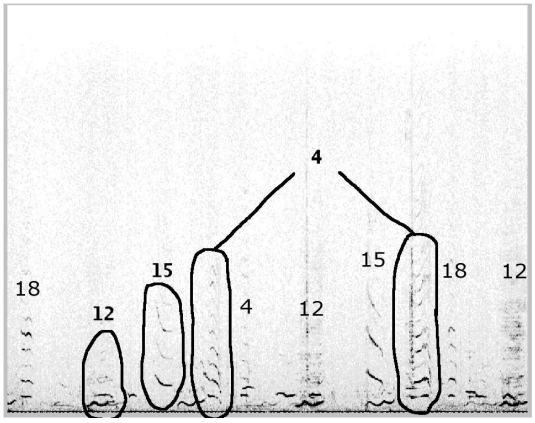

FIG. 10:

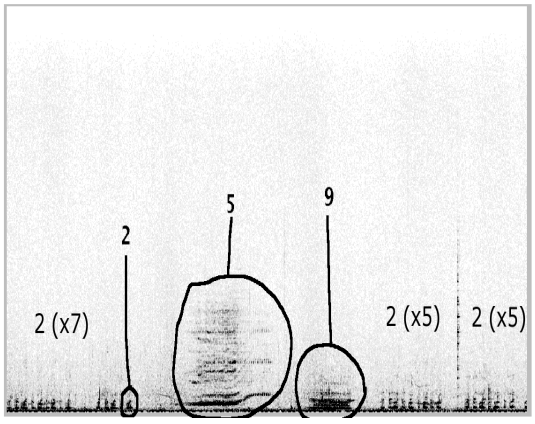

FIG. 11:

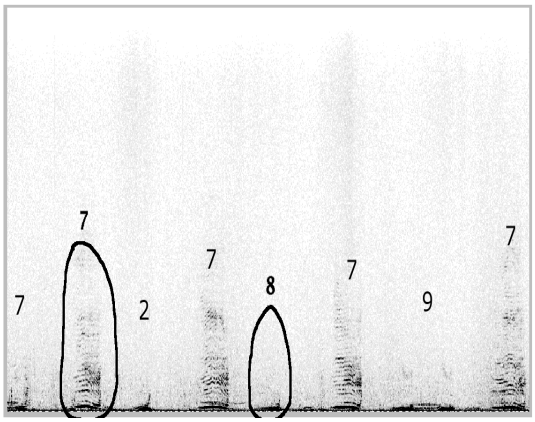

FIG. 12:

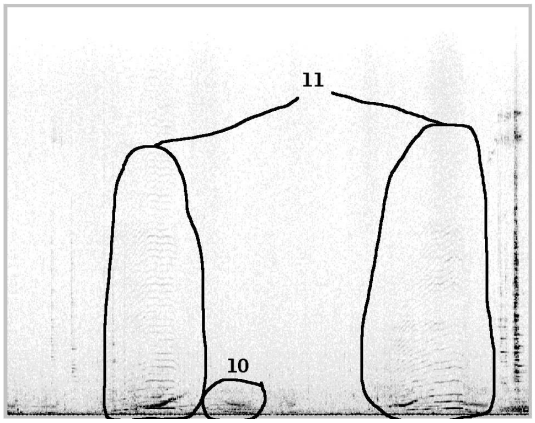

FIG. 15:

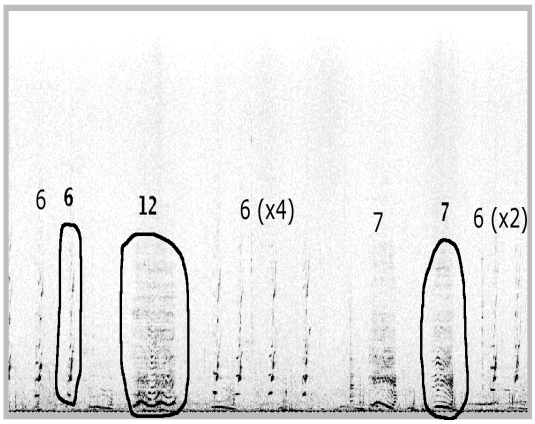

FIG. 13:

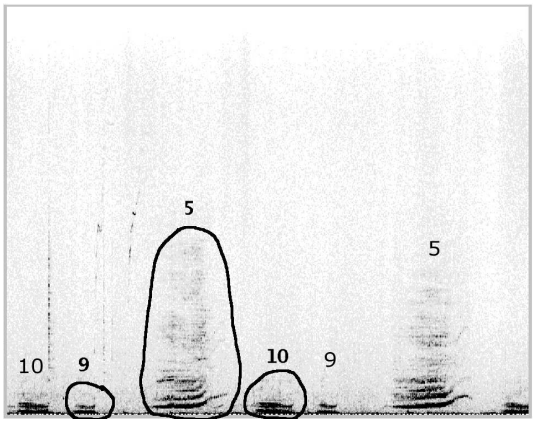

FIG. 14:

Stage	Parameter
Call Detection	FFT size and step
Call Detection	Signal frequency range
Call Detection	Detection segment size
Feature Extraction	VQ codebook size
Clustering	Number of initial categories
Clustering	HMM state duration constant

TABLE I: The complete list of user-supplied parameters required for end-to-end processing of raw sound data to clustered vocalizations.

VI. DISCUSSION

The good results achieved for humpback vocalizations demonstrate the power of these techniques. The vocalizations of humpbacks are particularly complex and they were all recorded under noisy real-world conditions. The technique succeeded in grouping similar calls together despite significant variations in signal levels, interference from other oceanic sounds, and recording equipment fluctuations.

The techniques require minimal manual configuration. Table I lists all user-specified parameters Essentially the users need only specified the broad character of the target vocalizations. Little effort was made to tailor the technique specifically for humpback vocalizations, and yet the achieved results were better than, or equal to previous specialized attempts in that domain ([6, 18]).

The simple approach described performs well, but we see many areas where the technique could be improved. The greatest improvement would be gained, we believe, by incorporating human feedback into the process. The segmentation procedure, for example, makes no attempt to use spectral features to differentiate noise and signal. While this has the advantage of keeping the technique simple and general, it does degrade performance. The ideal procedure, we believe, is to begin by making few assumptions about the spectral characteristics of the target signal, and then, after human input, rerun the algorithm with the benefit of this human expertise. For example, it is clear to any humpback expert (and indeed any non-expert who has spent some time listening to humpback calls) that 9 of the 19 'classes' produced by the classification algorithm are in fact ambient noise. If an expert can indicate this, and inform the program which 'classes' are just constituted of noise, the remaining classes, confirmed as containing actual humpback vocalizations, could be used to rerun the segmentation detection, but this time, both positive and negative examples (i.e. of noise and of signal) would be available. By looking for the distinguishing properties between the two, it should be possible to do substantially better than in the initial run, where no positive or negative examples are available. In effect, with only the brief intervention of an expert, it would be possible to 'bootstrap' the whole procedure from an unsupervised task to a supervised one, and yet this could be done with much less effort than in a wholly supervised procedure. The manual segmentation and classification of vocalizations by an expert would be reduced to the brief vetting of results produced by the initial unsupervised run.

Another area of improvement lies in allowing a human to vary the penalty function, producing either more, or fewer categories. We do not claim that the MDL-based penalty function employed produces the 'correct' number of clusters – indeed, we believe such a notion makes no sense. However, in this application, the technique does group calls in a way remarkably similar to human experts, who also typically recognise from around a half dozen to a dozen distinct types of call (see [5, 18, 28, 29, 42, 44]). The effect of the penalty function applied is to concentrate computational effort in looking for simpler models that fit the data. This is done not because simpler models are more probable (i.e. in the Bayesian prior sense), but because they are more preferable, since they are more comprehensible. If the aim is to maximize human intelligibility, then it makes sense for a human to have some input to the process. A human could do this quite simply by indicating if too many (or too few) classes were produced, and the penalty function could be adjusted accordingly.

VII. ACKNOWLEDGEMENTS

The Humpback recording were kindly supplied by Mike Noad (Uni of Qld.). He and Doug Cato (DSTO) discussed humpback vocalizations at length with us.

[1] A. Barron, J. Rissanen, B. Y. (**1998**). "The minimum description length principle in coding and modeling", IEEE Transactions on Information Theory **44**.

[2] Buck, J. and Tyack, P. (**1993**). "A quantitative measure of similarity for *Tursiops truncatus* signature whistles", Journal of the Acoustical Society of America **94**, 2497–2506.

[3] Burton, D. (**1987**). "Text-dependent speaker verification using vector quantization source coding", IEEE Transactions on Acoustics, Speech, and Signal Processing **35**, 133–143.

[4] Carpentera, G. and Grossberg, S. (**1987**). "ART 2: self-organization of stable category recognition codes for analog input patterns", Applied Optics **26**, 4919–4930.

[5] Cato, D. (**1991**). "Songs of humpback whales: the Australian perspective", Memoirs of the Queensland Museum **30**, 277–290.

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

[7] Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., and Freeman, D. (**1988**). "AutoClass: a bayesian classification system", in *Proceedings of the Fifth International Conference on Machine Learning*, 54–64 (Morgan Kaufmann Publishers).

[8] Das, S. and Picheny, M. (**1996**). "Issues in practical large vocabulary isolated word recognition: the IBM Tangora system", in *Automatic speech and speaker recognition*, edited by C. Lee, F. Soong, and K. Paliwal, 457–479 (Kluwer Academic Publishers).

[9] Dempster, A., Laird, N., and Rubin, D. (**1977**). "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society (B) **29**, 1–38.

[10] Domingos, P. (**1999**). "The role of Occam's Razor in knowledge discovery", Data Mining and Knowledge Discovery **3**, 409–425.

[11] Frazer, L. and Mercado, E. (**2000**). "A sonar model for humpback whale song", IEEE Journal of Oceanic Engineering **25**, 160–182.

[12] Frumhoff, P. (**1983**). "Aberrant songs of humpback whales (*Megaptera novaeangliae*): clues to the structure of humpback songs", in *Communication and Behavior of Whales*, edited by R. Payne, 81–128 (Westview Press).

[13] Furui, S. (**1981**). "Cepstral analysis technique for automatic speaker verification", IEEE Transactions on Acoustics, Speech, and Signal Processing **29**, 254–272.

[14] Furui, S. (**1986**). "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Transactions on Acoustics, Speech, and Signal Processing **34**, 52–59.

[15] Grigg, G., McCallum, H., Taylor, A., and Watson, G. (**1996**). "Monitoring frog communities: an application of machine learning", in *Eighth Innovative Applications of Artificial Intelligence Conference*.

[16] Grunwald, P. (**2004**). "A tutorial introduction to the minimum description length principle", in *Advances in Minimum Description Length: Theory and Applications* (MIT Press, Boston, MA.).

[17] Hafner, G., Hamilton, C., Steiner, W., Thompson, T., and Winn, H. (**1979**). "Signature information in the song of the humpback whale", Journal of the Acoustical Society of America **66**, 1–6.

[18] Helweg, D., Cato, D., Jenkins, P., Garrigue, C., and McCauley, R. (**1998**). "Geographic variation in south Pacific humpback whale songs", Behavior **135**, 1–27.

[19] Juang, B. (**1984**). "On the hidden markov model and dynamic time warping for speech recognition — a unified view", AT&T Bell Laboratories Technical Journal **63**, 1213–1238.

[20] Kogan, J. and Margoliash, D. (**1998**). "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: a comparative study", Journal of the Acoustical Society of America **103**, 2185–2196.

[21] Kohonen, T. (**1990**). "The self-organizing map", Proceedings of the IEEE **78**, 1464–1480.

[22] Lee, T. (**2001**). "An introduction to coding theory and the two-part minimum description length principle", International Statistical Review **69**, 169–183.

[23] Levinson, S. (**1985**). "Structural methods in automatic speech recognition", Proceedings of the IEEE **73**, 1625–1648.

[24] Makhoul, J., Roucos, S., and Gish, H. (**1985**). "Vector quantization in speech coding", Proceedings of the IEEE **73**, 1551–1558.

[25] Mercado, E. and Kuh, A. (**1998**). "Classification of humpback whale vocalizations using a self-organizing neural network", Proceedings of IJCNN'98 1584–1589.

[26] Oliver, J. and Baxter, R. (**1994**). "Mml and bayesianism: similarities and differences", Techical Report 206, Department of Computer Science, Monash University.

[27] Oliver, J. and Hand, D. (**1994**). "Introduction to minimum encoding inference", Techical Report 205, Department of Computer Science, Monash University.

[28] 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, 9–58 (Westview Press).

[29] Payne, R. and McVay, S. (**1971**). "Songs of humpback whales", Science **173**, 585–597.

[30] Potter, J., Mellinger, D., and Clark, C. (**1994**). "Marine mammal call discrimination using artificial neural networks", Journal of the Acoustical Society of America **96**, 1255–1262.

[31] Quinlan, J. (**1993**). *C4.5: Programs for Machine Learning* (Morgan Kaufmann).

[32] Rabiner, L. (**1989**). "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE **77**, 257–285.

[33] Rabiner, L. and Juang, B. (**1986**). "An introduction to hidden Markov models", IEEE ASSP Magazine 4–16, january edition.

[34] Rabiner, L., Juang, B., and Lee, C. (**1996**). "An overview of automatic speech recognition", in *Automatic speech and speaker recognition*, edited by C. Lee, F. Soong, and K. Paliwal, 1–30 (Kluwer Academic Publishers).

[35] Rabiner, L., Levinson, S., and Sondhi, M. (**1983**). "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition", The Bell System Technical Journal **62**, 1075–1106.

[36] Rabiner, L. and Sambur, M. (**1975**). "An algorithm for determining endpoints of isolated utterances", The Bell System Technical Journal **54**, 297–315.

[37] Rissanen, J. (**1987**). "Stochastic complexity", Journal of the Royal Statistical Society (B) **49**, 223–239.

[38] Rissanen, J. (**2001**). "Strong optimality of the normalized ml models as universal codes and information in data", IEEE Transactions on Information Theory **47**.

[39] Schaffer, C. (**1993**). "Overfitting avoidance as bias", Machine Learning **10**, 153–178.

[40] Shannon, C. (**1948**). "A mathematical theory of communication", Bell System Technical Journal **July/October**.

[41] Taylor, A. (**1995**). "Bird flight call discrimination using machine learning", Journal of the Acoustical Society of America **97**, 3370–3371.

[42] Thompson, P., Cummings, W., and Ha, S. (**1986**). "Sounds, source levels, and associated behavior of humpback whales, southest Alaska", Journal of the Acoustical Society of America **80**, 735–740.

[43] Wallace, C. and Boulton, D. (**1968**). "An information measure for classification", Computer Journal **11**, 185–194.

[44] Winn, H. and Winn, L. (**1978**). "The song of the humpback whale *Megaptera novaeangliae* in the West Indies", Marine Biology **47**, 97–114.