Proceedings of the Institute of Acoustics

STATISTICAL CHARACTERISATION AND CLASSIFICATION OF MARINE MAMMAL SOUNDS BY MULTIPLE-RESOLUTION ENCODING OF TRAINING DATA DISTRIBUTIONS

T J Hayward

Naval Research Laboratory, Washington, DC, USA

1. INTRODUCTION

Marine mammal vocalisations are of interest in bioacoustics both as aspects of echolocation and social behaviours and for their use in species identification for acoustic census of marine mammal populations. Many marine mammal sounds can be correctly classified to the species level, or at least to the level of a small number of related species, through audition by experienced marine mammal bioacousticians. This is often achieved in spite of substantial statistical variability of the received acoustic waveforms, resulting from both the variability of the source waveforms and the variability of the propagation characteristics of the ocean medium.

Recently, several research efforts have been directed at automated classification of marine mammal sounds, using hierarchical classifiers [1], artificial neural networks [2] and Bayesian classifiers [3]. For some vocalisation types, the statistical variability of the received waveforms presents a challenge to automated classification efforts. This is particularly true in cases where only a limited number of samples are available for use as training data in the preparation of an automated classifier.

Automated classification algorithms often are based on attempts to identify critical features for discriminating classes of vocalisations. In mathematical terminology, the classification is preceded by a mapping from the original space of observations (the data space) to a lower-dimensional *feature space* [4]. Although feature-based classifiers often have proven effective in practice, potential disadvantages of feature-based approaches may also be noted. In particular, there seems to be no universal method developed to date for *optimal* selection of a feature set, although optimisation may often be performed within a limited set of candidate features. Thus, classification performance may be limited by the choice of a particular feature map. In addition, features may be distorted by the effects of propagation in a complex, multipath ocean acoustic environment.

Reference [3] explores an alternative approach to classification, employing statistical characterisation of the training data distributions in the unreduced sample space. One potential advantage of this approach is that the statistical characterisation of a class may, in principle, be modified to include acoustic propagation effects for a prospective new site of acoustic reception. In addition, the classification algorithm in [3] makes effective use, through Bayesian inference, of what may be limited numbers of training samples.

This paper provides a brief description of the algorithmic approach developed in [3] and describes additional work to encode the statistical distributions of the training data more efficiently. Computational efficiency of the classification algorithm is assessed, and data storage requirements are compared with theoretical bounds.

2. CLASSIFICATION OF VOCALISATIONS

2.1 Supervised Learning Approach

The classification algorithm is based on supervised learning; that is, the training of the automated classifier is based on the classification decisions of a human expert, applied to training data consisting of recorded vocalisations. These classification decisions should be based on visual confirmation of the species of the vocalising animal [1,5].

Formally, the training data are structured and utilised as follows: The expert groups the training samples into K sets $\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_K$, with $\mathcal{T}_k$ denoting the set of N_k training samples

$$\mathcal{T}_k = \left\{ S_1^{(k)}, S_2^{(k)}, ..., S_{N_k}^{(k)} \right\}, \qquad k = 1, 2, ..., K. \tag{1}$$

Each training set may, for example, consist of vocalisation samples from a particular species. Each sample consists of a received pressure waveform or transformed data, such as a sonogram. In the classification phase, a new sample S is presented to be classified as belonging to one of K 'classes' presumed to be associated with the training sets.

2.2 Bayesian Classification

The automated classification decision is based on computation of *a posteriori* probabilities of membership of the new sample in each class. These computations are based on Bayesian inference, applied to successive partitions of the sample space Σ. (Σ is the set of all possible samples of the type under consideration.) Details of this partitioning process are given in [3]; Section 3 (below) provides an outline. At each stage of the process, and for each class, we determine the number of training samples in that class that belong to the same cell of the partition as does the new sample. These integer-valued statistics are used recursively, as the partition is refined, to compute the *a posteriori* probabilities.

2.3 Classification Performance

An initial demonstration of the classification algorithm is described in [3]. Training data consisted of 64 sonograms, representing six classes of low-frequency baleen whale sounds. An overall rate of 75% correct classifications was achieved despite substantial statistical variability within several of the classes and despite the small number of training samples used.

2.4 Computational Efficiency

The algorithm operates directly on the data bits (binary digits) of the training samples and the samples to be classified. Apart from the signal processing (e.g., sonogram computation), very few arithmetic operations are performed. This leads to a highly efficient computation for both training and classification. For the example cited above, the computation time, including training data file access, classification and output, but not including the sonogram computation, was approximately 330 ms per sample (vocalisation) on an SGI Indy 5000 workstation. This is approximately one-fourth the average sample duration of 1.3 s. This high speed is achieved through a highly efficient binary encoding of the samples.

3. ENCODING OF TRAINING DATA

3.1 The Sample Space

Samples presented for classification may consist of unprocessed time series or, more usually, any of several standard transforms of time series, such as Fourier transforms, time-frequency distributions (including conventional sonograms), wavelet transforms, etc. In any of these cases, a sample S may be described mathematically as a subset of the Cartesian product X of d bounded real-number intervals; i.e.,

$$S \subseteq X = X_1 \times X_2 \times \cdots \times X_d, \tag{2}$$

where X_i denotes a real interval $[x^{(i)}_{\min}, x^{(i)}_{\max})$. For example, an acoustic pressure time series consists of a set of data points

$$S = \{(t_1, x_1), (t_2, x_2), \ldots, (t_n, x_n)\}, \tag{3}$$

where t_n is a sampling time and x_n is the received acoustic pressure at that time. In that case, $d = 2$, and S is a subset of the Cartesian product of a time interval $[t^{(i)}_{\min}, t^{(i)}_{\max})$ and an interval $[x^{(i)}_{\min}, x^{(i)}_{\max})$ of acoustic pressures. Similarly, a sonogram consists of a subset of the Cartesian product of a time interval $[t^{(i)}_{\min}, t^{(i)}_{\max})$, a frequency interval $[f^{(i)}_{\min}, f^{(i)}_{\max})$, and a power interval $[p^{(i)}_{\min}, p^{(i)}_{\max})$. In either case, the sample space Σ consists of the set of all possible samples, i.e., the set of all subsets of X.

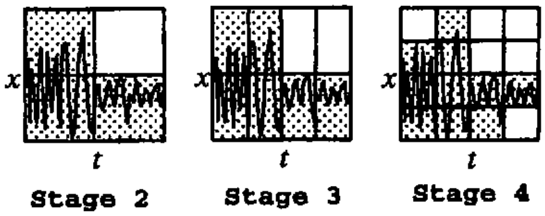

Figure 1: Three stages of the multiple-resolution partition, illustrated schematically for a time series.

3.2 Multiple-Resolution Partition of the Sample Space and Training Data Histograms

An individual sample may be partitioned in binary fashion as indicated (for a time series) in Figure 1. Details of this partitioning process are provided in [3]. This process provides an efficient binary encoding of the training samples. In turn, the partition of samples induces a multiple-resolution partition of the sample space Σ. At the λ-th stage of this partition, Σ is divided into 2^λ cells numbered $\Sigma_{b_1 b_2 \cdots b_\lambda}$, where each b_i takes the value 0 or 1. The training data may then be statistically characterised by counting the number of elements of each training set $\mathcal{T}_k$ that belong to each cell $\Sigma_{b_1 b_2 \cdots b_\lambda}$; denote this number by $\nu^{(k)}_{(b_1 b_2 \cdots b_\lambda)}$. These numbers constitute (high-dimensional) histograms of the training data in the sample space. Figure 2(a) illustrates, schematically, the four cells at the third stage of this partitioning process, and the occupancy of these cells by training samples for three classes. Figure 2(b) illustrates the corresponding histogram. The histograms provide a detailed, nonparametric description of the statistical distribution of the training data.

3.3 Efficient Computer Storage of Training Samples

The partitioning process terminates when the resolution limit of the available data is reached; however, a lower resolution may be chosen to improve computational speed if it is found to be adequate for classification. The storage requirement, in bits, for training samples encoded to a maximum resolution of $2^{\lambda_{\max}}$ cells is given by

$$\sum_{\lambda=1}^{\lambda_{\max}} \sum_{\substack{b_1,b_2,\ldots,b_\lambda=0,1; \\ v^{(k)}_{(b_1b_2\ldots b_\lambda)}\neq 0}} \lambda v^{(k)}_{(b_1b_2\ldots b_\lambda)}. \qquad (4)$$

The choice of a relatively low maximum resolution may permit substantial compression of the training data without loss of class discrimination capability. For example, for the six-way classification described above, a data compression ratio of 10.7:1 relative to the original time series was achieved, without loss of classification performance, by using $\lambda_{\max}$=12. This corresponds to a partition of the sonograms having 16 cubes on each side (time, frequency, and power).

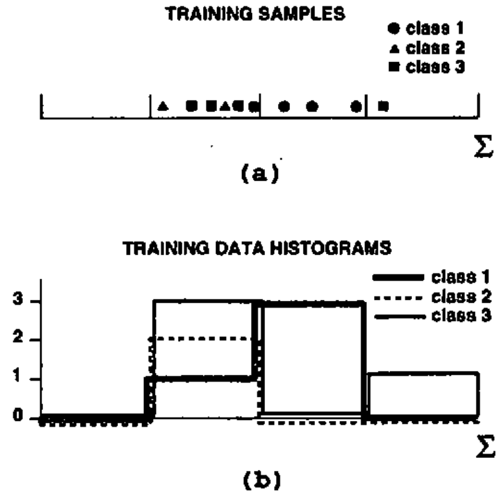

Figure 2: Schematic illustration of (a) training data for three classes, and partition of the sample space Σ; and (b) the corresponding training data histograms.

3.4 Efficient Computer Storage of Training Data Histograms

The training data may be further compressed by storing training data histograms rather than the individual encoded samples. The number of bits that are required to store the training data histograms for the *k*-th class, up to a maximum resolution of $2^{\lambda_{\max}}$ cells in the partition of each sample, is given by

$$\sum_{\lambda=1}^{\lambda_{\max}} \sum_{\substack{b_1, b_2, \ldots, b_\lambda = 0,1; \\ \nu^{(k)}_{(b_1 b_2 \ldots b_\lambda)} \neq 0}} \left(\log_2 \nu^{(k)}_{(b_1 b_2 \ldots b_\lambda)} + \lambda \right). \tag{5}$$

This is simply the number of bits required to specify each occupied cell in the sample space and the number of samples occupying that cell. Thus, substantial storage can be saved by storing the training data histograms rather than the individual encoded training samples. For the six-way classification example, Eq. (5), with $\lambda_{\max}$= 12, provides a data compression ratio of 3938:1 compared to the time series. A more computationally efficient, and readily implemented, storage scheme allocates 16 bits (two 8-byte words) for each number $\nu^{(k)}_{(b_1 b_2 \cdots b_\lambda)}$; this provides a data compression ratio of 2108:1 for the six-way classification example.

For comparison, we note that the theoretical bound on the storage requirement for encoding the training data, determined by information theory [6], is given by

$$- \sum_{\substack{b_1, b_2, \ldots, b_\lambda = 0,1; \\ \nu^{(k)}_{(b_1 b_2 \cdots b_\lambda)} \neq 0}} \nu^{(k)}_{(b_1 b_2 \cdots b_\lambda)} \log_2 \left(\frac{\nu^{(k)}_{(b_1 b_2 \cdots b_\lambda)}}{N_k} \right), \tag{6}$$

with $\lambda = \lambda_{\max}$. For the six-way classification example, Eq. (6) provides a possible data compression ratio of 9280:1 compared to the time series. However, implementation would require additional computation time for the training and classification.

4. SUMMARY AND DISCUSSION

An efficient classification and training-data compression algorithm has been developed and implemented, and its potential for automated classification of marine mammal vocalisations has been investigated. The computation speeds demonstrated so far show promise for a real-time classification capability. Simultaneously, very large (over 2000:1) data compression ratios for the training data were achieved without loss of computational efficiency or classification performance.

An unusual feature of the algorithm is its lack of reliance on feature identification. Because the classification is based on statistical characterisation of training data in the unreduced sample space, it may be possible to adapt training data to new underwater environments based on a statistical representation of local propagation effects. In addition, the sample-space histograms provide a suitable structure for determination of training data requirements, based on measures of statistical dispersion within each class.

In [3], a rate of correct classification of 75% was achieved for six classes of baleen whale sounds, based on a total of 64 training samples. While this is good performance in view of the signal variability and the small training sets, there is room for improvement. A potential strength or weakness of the current algorithm is its canonical progression from large to small-scale signal components. This progression may be advantageous in many cases, since classes may often be discriminated first by their large-scale characteristics, but one can envision cases where it may not be optimal. The optimisation of the sample-space partition will be investigated in future work.

A further question is whether any additional data compression may be achieved by simultaneous (e.g., hierarchical) encoding of the training data for several classes. Currently, the training data distributions for each class are encoded separately. Simultaneous encoding may be advantageous for large numbers of classes, as suggested by the results in [1].

The algorithm may be applied to any signal processor output. This leaves open the question of what signal transform should be used, i.e., can the processing be chosen to optimise classification performance, and can the partition be optimised for the processing?

5. ACKNOWLEDGEMENT

This work was supported by Office of Naval Research Base Funding at the Naval Research Laboratory, Washington, DC, USA.

6. REFERENCES

[1] K M FRISTRUP & W A WATKINS, 'Marine animal sound classification,' Technical Report WHOI-94-13, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, USA (1994)

[2] J R POTTER, D K MELLINGER & C W CLARK, 'Marine mammal call discrimination using artificial neural networks,' *J Acoust Soc Am*, **96** p1255 (1994)

[3] T J HAYWARD, 'Classification by multiple-resolution statistical analysis with application to automated recognition of marine mammal sounds,' *J Acoust Soc Am*, **101** p1516 (1997)

[4] R J SCHALKOFF, *Pattern Recognition: Statistical, Structural, and Neural Approaches*, John Wiley and Sons, New York (1992)

[5] W A WATKINS, K M FRISTRUP, M A DAHER & T HOWALD, '*SOUND* Database of Marine Animal Vocalizations; Structure and Operations,' Technical Report WHOI-92-31, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, USA (1992)

[6] C E SHANNON & W WEAVER, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, Illinois, USA (1949)