Classification of Transient Sonar Sounds Using Perceptually Motivated Features

Simon Tucker and Guy J. Brown

Abstract—**This paper describes a novel framework for classifying underwater transient signals recorded by passive sonar. The proposed approach involves two key ideas. Firstly, a feature-selection algorithm is used to identify those acoustic features that optimally model each class of transient sound. Secondly, features that are perceptually motivated are proposed, i.e., they encode information that human listeners are likely to use in transient classification tasks. Three perceptual features are proposed, which encode timbre, the physical material of the sound source, and the temporal context (pattern) in which the transient occurred. The authors show how these features, which are computed over different temporal windows, can be combined to make classification decisions. The performance of the proposed classifier is evaluated on a corpus of transient signals extracted from passive sonar recordings. Specifically, the performance of the perceptual features is compared with spectral features and with those that encode statistics of time, frequency, and power. The present results show that the perceptual features provide valuable cues to the class of a transient. However, the best performing classifier was obtained by selecting a subset of perceptual, spectral, and statistical features in a class-dependent manner.**

Index Terms—**Auditory model, passive sonar, transient analysis, transient classification.**

I. Introduction

THE classification of underwater signals recorded from passive sonar remains a challenging problem, which is of immense interest for military applications. Conventionally, human experts perform identification by listening, or by visual inspection of spectrographic representations. However, the increasing complexity of sonar arrays means that manual classification is often impractical; there is therefore a real need for reliable automatic classification.

Long-duration underwater sound emissions such as engine noise can be largely suppressed, whereas transient emissions (such as those caused by a buckling hull or propeller cavitation) are harder to control. Accordingly, there has been particular interest in the automatic classification of transient sonar sounds (e.g., [3], [16], [18], [22], [23], [28]). However, reliable classification is difficult because such signals vary widely in their temporal and spectral characteristics and can originate from biological sources (such as shrimp and cetacea) as well as mechanical ones. Furthermore, transients of interest are often of very short duration and tend to be embedded in high levels of ambient ocean noise.

Various approaches to automatic classification of sonar transients have been described, including those based on time–frequency analysis [3], [22], fuzzy logic [18], and a hybrid hidden Markov model-multilayer perceptron (HMM-MLP) classifier [16]. However, the superior performance of experienced human listeners in transient classification tasks suggests that some advantage might be gained from a perceptually motivated approach. There is some support for this notion in the literature. For example, Teolis and Shamma [28] found that spectral features derived from a computational auditory model gave better performance in a sonar classification task than spectra obtained from the short-time Fourier transform (STFT). Similarly, Parks and Weisburn [23] report that a perceptually motivated frequency analysis offers an advantage over the STFT for the classification of whale and ice sounds.

Accordingly, this paper describes a novel approach to the classification of sonar transients which is motivated by two underlying principles. Firstly, there is a focus on acoustic features that are held to be important for human classification of transient sounds, as determined by psycho-physical experiments. Secondly, the authors deal with the problem of varying temporal and spectral characteristics by using a feature selection algorithm, which determines a subset of features that optimally model each class of transient sounds.

Three kinds of acoustic features are employed in the current study; those which provide cues to the timbre of a transient sound, cues to the material properties of its underlying physical source, and an encoding of its temporal context (see Fig. 1). Clearly, information from these features must be integrated in order to make a classification decision, and this raises the issue of how to combine features that are extracted at different time scales. For example, consider a sequence of related transient events (such as propeller cavitation). The timbre of each event might be described using features computed on a millisecond time scale, whereas the signal may also contain a temporal pattern that extends over several seconds. Here, it is shown how features computed at different time scales can be effectively combined by embedding information about long-duration temporal patterns into feature vectors computed over a short time period.

Comparative evaluation of transient classification systems is somewhat problematic, because there is no widely available standard corpus, and much of the data used by other workers cannot be distributed for reasons of national security. Here, the authors' own data set is used, but the performance of the proposed classification system is compared across four

Manuscript received July 16, 2004; accepted February 11, 2005. The work of S. Tucker was supported by an EPSRC CASE Award in association with the Sound Concepts Department, QinetiQ, Winfrith. The work of G. J. Brown was supported by EPSRC under Grant GR/R47400/01.

S. Tucker is with the Department of Information Studies, University of Sheffield, Sheffield S1 4DP, U.K. (e-mail: s.tucker@shef.ac.uk).

G. J. Brown is with the Department of Computer Science, University of Sheffield, Sheffield S1 4DP, U.K. (e-mail: g.brown@dcs.shef.ac.uk).

Digital Object Identifier 10.1109/JOE.2005.850910

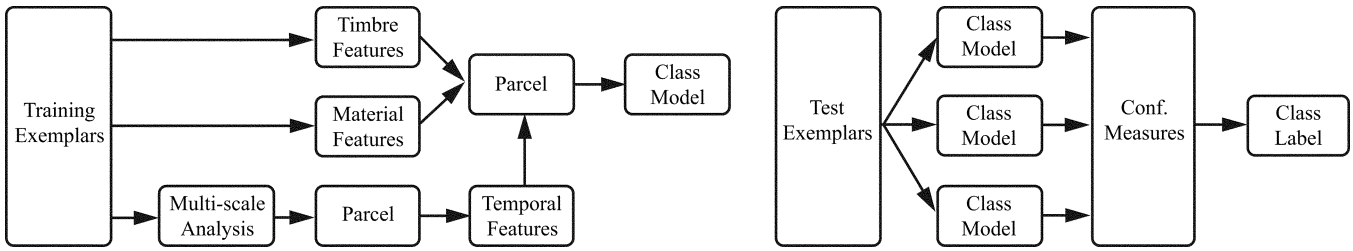

Fig. 1. Overview of the training (left panel) and testing (right panel) procedures. During training, three kinds of features are computed for each exemplar in the training set. The Parcel algorithm, which is a means of selecting an effective subset of features based on classification performance, is used twice during training: once to select temporal features at an appropriate time scale, and then again to determine the features appropriate for modeling each class in the corpus. During testing, the feature vectors for each transient class are computed, and a classifier is used to determine the confidence value for each class model. The class with the highest confidence value is taken as the label of the test exemplar.

conditions using the following: 1) perceptually motivated features; 2) spectral features derived from the STFT; 3) features that measure the statistics of time, frequency, and power; and 4) perceptual, spectral, and statistical features together. It is demonstrated that perceptually motivated features are often selected in preference to the others, but that the best performing classifier uses a subset of all the available features.

The rest of the paper is organized as follows. Section II gives an overview of the procedure used to train and test the transient classifier. Section III describes the corpus of transient sonar sounds that have been used for evaluating the proposed system. In Section IV, the perceptually motivated acoustic features are described, and a means of integrating different features is outlined. The listening experiments that were carried out to derive the perceptually motivated features are only briefly described, because the focus of the current paper is on the classifier itself rather than the underlying psychophysical justification for it; for further details, the reader is referred to [32]. Section V describes how the optimal feature set is derived for each class, and Section VI explains the classification algorithm. The experimental protocol for evaluating the classifier is described in Section VII, and results are presented in Section VIII. Section IX concludes with a summary and discussion.

II. Overview of the System

The procedures for training and testing the transient classifier are shown in Fig. 1. In the training phase, acoustic features are computed for each exemplar in the training set. Features that encode the temporal pattern of events are computed at a number of time scales, forming a hierarchical view of the temporal context around each transient event. The Parcel algorithm [27] is then used to perform a heuristic search in order to find the subset of temporal features for each class, which maximally discriminates it from the remaining classes. The selected temporal features are supplemented by features that encode the timbre and perceived material of the individual transient event. Parcel is then used again to select the optimal subset of features for each class from all of those available.

It should be noted that Parcel evaluates a number of classifiers for each class (we use the term "classifier" to denote the combination of a classification algorithm and a particular set of features). Each classifier has a different operating point (i.e., a different tradeoff between true positives and false positives). Consider two classes of transient, i and j. A true positive occurs when a transient belonging to class i is correctly assigned to class i, and a false positive occurs when a transient belonging to class j is erroneously assigned to class i. Here, the tradeoff between true positives and false positives is determined by the feature set used, and by tuning a parameter of a fuzzy k-nearest neighbors (KNN) classifier. The ability to specify an operating point is highly desirable, because transient detection is inherently a variable-cost domain. For example, in military applications, the cost of a misclassification may depend on whether transient detection is being carried out during routine ocean monitoring, or during combat.

In the testing phase, an operating point is specified, and the corresponding feature set and fuzzy KNN tuning parameter are retrieved for each class model. Test exemplars are then matched against each class model, and labeled with the class that yields the highest confidence value (for a justification of this approach, see [14]).

The following sections describe the training and testing procedures in detail. First, the corpus of transient sonar sounds that were employed is described. Then, the acoustic features that were used are explained, and methods for combining features that cover different temporal contexts are suggested. Finally, the feature-selection process and fuzzy KNN classifier are described.

III. Corpus

The corpus used for evaluation consisted of a collection of passive sonar recordings of biological and mechanical sources, recorded at a number of different ocean locations. The acoustic signals were digitally sampled at a rate of 24 kHz with a 16-bit resolution, and they were supplied to us with ground-truth labels provided by experienced sonar operators.

The start and end points of transient events were manually identified. For each recording, noise reduction was performed using spectral subtraction [4]. Specifically, the spectrum of the ambient ocean noise was estimated during a period in which there was no transient activity, and this was subtracted from the spectrum of each noisy signal. Overall, 1200 transients were identified from 10 different transient classes. These data were split into training, validation, and testing sets in the ratio of 2:1:1. The transient classes selected and the number of exemplars in each class are shown in Table I. Note that the signals

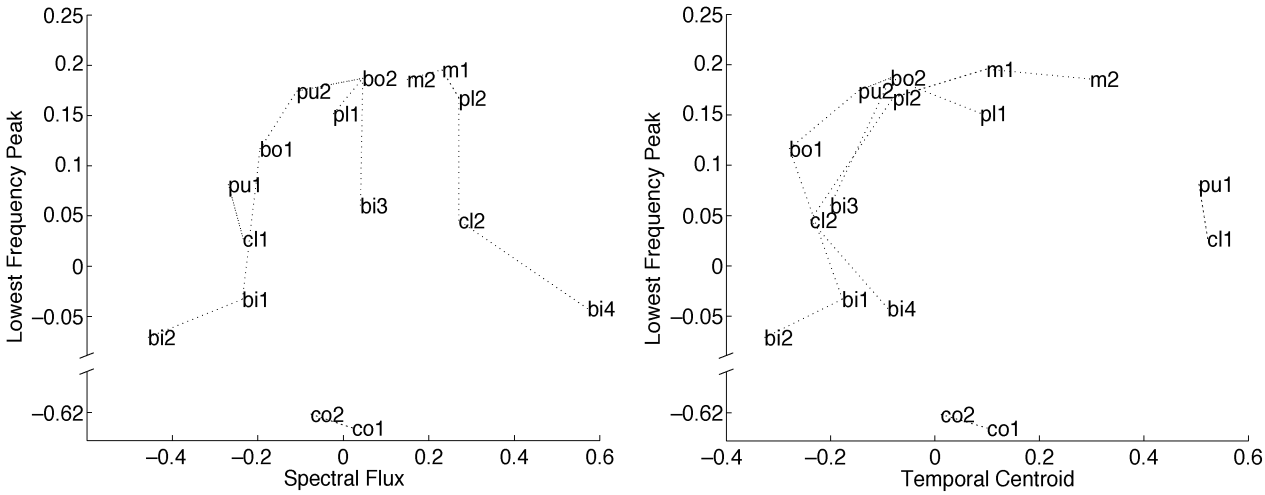

Fig. 2. Two-dimensional projections of timbre space constructed from listener judgements of the similarity of sonar sounds. BI = biological sounds, BO = bow movements, CL = clanks, CO = counter signals, M = mast movement, PL = plane movement and PU = pump sounds. Neighbors are connected by dotted lines. The counter signals were characterized by a narrow bandwidth, which explains their separation from the other sounds along the "lowest frequency peak" dimension.

TABLE I
CLASSES OF TRANSIENT IN THE CORPUS. THE NUMBER OF EXEMPLARS IN EACH CLASS AND THE MEAN DURATION OF THE SIGNALS IN EACH CLASS ARE SHOWN

Class	Code	Count	Mean duration (millisecond)
Bio Clicks	BC	201	197
Mooring	MO	43	180
Propeller	PR	245	140
Controls	CO	118	353
Rattles	RA	95	161
Chains	CH	58	413
Machinery	MA	40	90
Knocks	KN	41	217
Tinkling	TI	314	104
Propulsion	PN	45	112

differed substantially in duration; mean durations for each class varied from 90 to 413 ms.

IV. ACOUSTIC FEATURES

A. Timbre

Timbre is defined by the American National Standards Institute (ANSI) as being "that attribute of auditory sensation in terms of which a subject can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" [1]. This definition is somewhat problematic for transient sounds, which being of short duration often elicit no clear pitch percept—nonetheless, they have a distinct timbre. A more appropriate definition for these purposes is given by Handel [13], who simply defines timbre as " . . .what it sounds like."

Timbre is widely regarded as a multidimensional percept, and psychophysical studies of timbre typically employ multidimensional scaling (MDS) algorithms to identify the perceptual dimensions [12]. Such studies collect listener judgements of similarity for a small number of tones and transform the judgements into a space of two to four dimensions. A transformation is chosen such that sounds that were consistently judged as being similar are located close to each other in the timbre space, whereas those that were judged to be dissimilar are placed far apart. The acoustic cue corresponding to each dimension of the space is then identified by correlating the acoustic properties of the sounds with their corresponding location along each dimension of the timbre space.

A similar approach was used here. Sixteen transients of both biological and mechanical origin were manually selected from passive sonar recordings. Ten subjects judged the similarity of pairs of transients on a five-point scale, ranging from "similar" to "dissimilar." These judgements were transformed to a three-dimensional (3-D) timbre space using the INDSCAL algorithm [6]. Two-dimensional (2-D) projections of the timbre space are shown in Fig. 2.

Numerous acoustic features were then computed for the transient signals, including those previously described in the literature on musical timbre (e.g., see [21] and [24]) and various features derived from STFT spectra. The correlation between each feature and the dimensional location of each sound was determined using the Pearson product moment correlation coefficient, and the corresponding acoustic feature was chosen, which gave the highest correlation for each dimension.

Three acoustic features were identified by this process, which correspond respectively to the three dimensions in Fig. 2. Note that to produce the acoustic features described below, the signals were split into frames of 10-ms duration with a 5-ms overlap. The first was a measure of the spectral change over time (denoted "spectral flux" by [21]), which was computed as the mean correlation between successive spectra derived from the STFT

$$\mathrm{SF} = \frac{1}{M}\sum_{k=1}^{M} |r_{k,k-1}|. \tag{1}$$

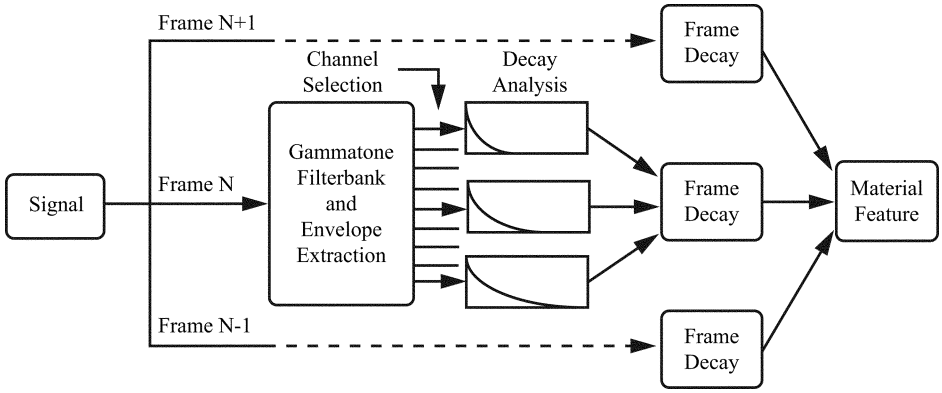

Fig. 3. Schematic diagram of the algorithm for determining the material feature. Note that all time frames contribute to the estimation of the material property, but for clarity, only three frames are shown here.

Here, M is the number of frames in the signal and $r_{k,k-1}$ is the Pearson product moment correlation coefficient between the vectors of spectral features at time frames k and $k-1$. The second feature was the frequency of the lowest peak in the average spectrum (computed as the mean spectrum over all time frames), with the constraint that the amplitude of this peak should be no less than 10 dB below the maximum spectral amplitude.

The third feature identified was the temporal centroid of the transient; in other words, the center of gravity of its temporal envelope. This was computed by

$$\mathrm{TC} = \frac{\sum_t tE(t)}{\sum_t E(t)} \tag{2}$$

where $E(t)$ is the signal energy at time t (see also [24]). Here, it is assumed that transients have been detected and isolated by a prior process (see Section VII), and hence, $t = 0$ was taken to be the starting time of the isolated transient.

B. Material

Numerous studies have investigated the perception of physical properties of vibrating objects, such as the hardness of mallets [10], the gender of walking sounds [19], and acoustic correlates of the shape [17] and material [11] of vibrating plates. However, little consideration has been given to the role that the acoustic environment plays in such listening tasks (although [11] examines the effect of external damping on the perception of material properties). It is therefore unclear whether the acoustic cues available to listeners when sound is propagated in air are also available when sound is propagated in water.

Accordingly, the perception of physical properties of objects vibrating both in air and underwater was investigated. Recordings were made by striking plates of a variety of sizes, shapes, and materials when they were suspended in an acoustic booth and in a large water tank. These recordings formed the basis for a psychophysical experiment that asked subjects to estimate the size ratio of two plates of the same shape and material and also asked subjects to estimate the shape and material of plates of the same size (for details see [33]).

The results of this study showed that subjects were very good at identifying the material of the plates, and also that their judgements formed characteristic "macrocategories" (see also [11]). Macrocategories are groups of stimuli for which confusion within a group is significantly higher than confusion between groups. In this experiment, it was found that metallic plates were placed in one macrocategory and that another was formed from plastic and wood plates. Furthermore, comparison of listener's performance for in-air and underwater recordings suggested that the rate of decay of transient sounds was an important factor in determining listener's perception of material properties.

Motivated by these experimental findings, a computer model that extracts a cue to material property from acoustic transients was developed. Classification experiments using this model suggest that it gives a good match to listener's performance [33]. The stages of processing in the computer model are shown schematically in Fig. 3.

The acoustic impulse response of a freely vibrating plate can be described as the superposition of a number of decaying sinusoids, the frequencies of which are determined by the physical properties of the plate itself. Each sinusoid $p(t)$ has the form

$$p(t) = a\mathrm{e}^{-tf\pi\tan\phi}\sin(2\pi ft) \tag{3}$$

where f is the frequency of the component, t is the time, a is the initial amplitude, and $\tan\phi$ is a material-dependent parameter referred to as the internal friction [9]. Hence, the decay of each frequency component depends only on the material properties of the plate, and may be regarded as a reliable indicator of material.

Accordingly, processing in the computer model proceeds as follows. Initially, the sampled signal is filtered by a bank of 128 bandpass (gammatone) filters, which model the frequency selectivity of the cochlea. Filter center frequencies are spaced between 50 Hz and 12 kHz on an ERB-rate scale [5]. The Hilbert envelope is computed in each channel and integrated over a 200-ms window with a 10-ms overlap. Spectral peaks are then identified by convolving each short-term spectrum with the derivative of a Gaussian and locating the zero crossings. Only frequency channels that correspond to spectral peaks are selected for further processing. Specifically, the decay in the selected channels is estimated using the Prony method [20], and the median decay across all channels is taken as the estimate of $\tan\phi$ for that time frame. The mean value over all time frames,

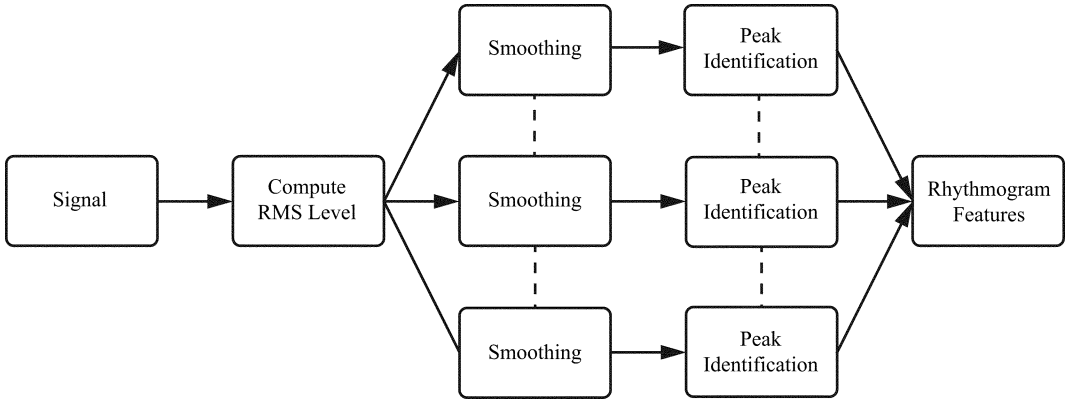

Fig. 4. Computation of rhythmogram features. The rms level of the signal is computed and smoothed at various time scales. Peaks in amplitude are identified at each smoothing scale, and they are combined to provide features for classification.

weighted by the energy in each frame, is then computed to give a material value for the overall signal.

C. Temporal Pattern

Many of the sonar transients considered here do not occur as isolated events. For example, propeller cavitation consists of a sequence of transient sounds that occur in a regular temporal pattern; the "popping" produced by snapping shrimp is another example. Analysis of temporal pattern can therefore provide a useful basis for the classification of certain sonar sounds.

Todd [29] describes an auditory-motivated approach to analyzing temporal structure in acoustic signals, called the rhythmogram. The rhythmogram is a multiscale analysis, which may be regarded as an auditory analog to edge detection in vision. To construct the rhythmogram, the temporal envelope of a signal is progressively low-pass filtered and peaks are identified at each level of smoothing. By plotting the temporal location of peaks against the degree of smoothing, a hierarchical view of temporal structure is developed. Todd demonstrates that the rhythmogram is able to identify the hierarchical temporal structure present in music and speech [30].

Here, a representation of temporal pattern, which is similar to Todd's, is computed; however, the motivation is different since features for classification, rather than a visual display for acoustic analysis, are required. The procedure used is shown schematically in Fig. 4. First, the temporal envelope of the signal is extracted by computing the root-mean-square (rms) level over a window of size 30 ms. Gaussian filters are then used to smooth the envelope at a number of scales, varying from 0.2 to 122 ms. Peaks are identified in the smoothed envelope at each scale to give a hierarchical representation, an example of which is shown in Fig. 5.

The representation shown in the figure is not suitable as a basis for classification for two reasons. Firstly, the rhythmogram is time dependent, i.e., a different structure emerges depending on the time at which the temporal sequence starts. Secondly, the rhythmogram contains redundant information, because a range of smoothing filters detect the same temporal events. This is apparent in Fig. 5, in which the structure between Gaussian widths of 20 and 60 ms is nearly identical.

The time dependency of the rhythmogram can be addressed by performing classification upon the distribution of interpeak

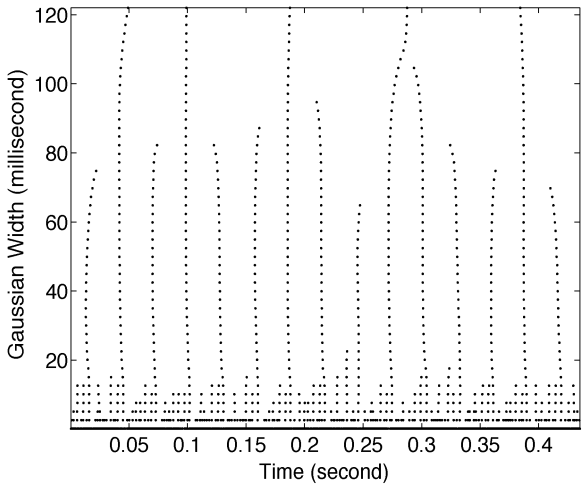

Fig. 5. Rhythmogram representation of a repeating pattern of knocks. The temporal pattern is evident from the hierarchical structure of the rhythmogram; five events, each of which consists of a sequence of between two and four transients.

intervals at each time scale, rather than the absolute times at which peaks occur. Specifically, interpeak interval histograms are computed at each scale, which are normalized by the number of peaks. Normalization removes a scale-dependent bias in the frequency of events; it is apparent from Fig. 5 that the frequency of detected events is inversely proportional to the width of the smoothing filter. In practice, since the frequency bins are sparsely populated, it was found that it was preferable to use a parametric approach in which the distribution of interpeak intervals at each scale was represented by a mixture of Gaussians (MOG). The expectation maximization algorithm [2] was used to fit a MOG to the interpeak interval histograms, as shown in Fig. 6.

The problem of redundancy in the rhythmogram might be addressed in a number of ways. A simple approach is to reduce the rhythmogram to a single vector by averaging over the scale dimension. However, this approach did not perform well, presumably because information about the activity at different scales is lost. Better classification performance was obtained by using a representative subset of scales, but selecting the appropriate scales proved to be nontrivial. Attempts were made to choose a subset of scales according to a linear division of scale space

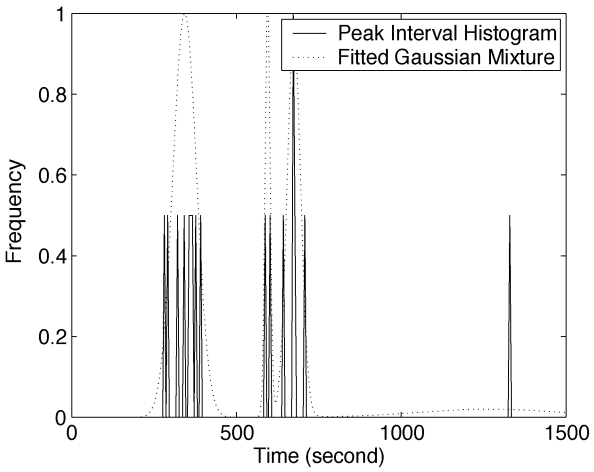

Fig. 6. Example of an interpeak interval histogram derived from the rhythmogram (solid line) and the corresponding mixture of Gaussians fit (dotted line).

and by selecting scales at which the most activity was present. Neither approach was satisfactory.

A solution to the problem of scale selection from the rhythmogram eventually came from two insights. Firstly, the most appropriate set of scales depends on the class of transient that is being modeled, since sonar sounds differ widely in their temporal structure. Accordingly, scale selection is done independently for each class in the training set. Secondly, scale selection is essentially a kind of feature selection; a heuristic search of scale subsets is therefore performed using the Parcel algorithm (see Section VII and [27]).

During training, features for encoding temporal pattern are computed and selected for each class as described above. The temporal context used to compute the rhythmogram is a 10-s window centered on the transient even. During testing on novel stimuli, the rhythmogram is computed, and the likelihood of a class model producing the observed temporal pattern is derived from the MOG for each scale. Specifically, the sum of the probabilities derived from each mixture component of the MOG is used as a measure of scale likelihood. To prevent bias between models with a different number of scales, the mean of the scale likelihoods is computed and used as an overall class likelihood. A feature vector is then constructed by concatenating the likelihoods for each class. In other words, a vector of temporal features $\mathrm{TE} = \{c_1, c_2, \ldots, c_N\}$ is derived, where c_i is the likelihood that class i produced the observed temporal pattern and N is the number of transient classes.

D. Interim Summary

In summary, transient sonar sounds are represented using three kinds of features. Firstly, measures related to the timbre of the sound are used, which encode the frequency of the lowest spectral peak, the spectral flux, and the temporal centroid. Secondly, features that cue the perceived material of the sound are employed; these encode the rate of decay of significant spectral components. Finally, temporal pattern is encoded using a multiscale analysis of changes in the envelope of the acoustic signal. The following two sections describe how a classifier can be constructed based on these features.

V. Feature Selection

This section describes how acoustic features that optimally represent each class of transient sound are selected. Three distinct sets of signals are involved in this process. The training set is used to train acoustical models for each class, and these are subsequently evaluated on a validation set during the feature-selection process. Once the model parameters have been determined and appropriate features have been selected for each class, the classifier is evaluated on a test set.

A. Determining Optimal Feature Vectors

As noted previously, transients differ widely in their spectral and temporal characteristics, and therefore a class of transient may be optimally modeled by a subset of the available acoustic features. For example, several classes of transient in this training corpus occur only as isolated events, and not in a sustained temporal pattern. Features that encode temporal pattern will not aid classification in such cases, and may even impair it.

The search for a subset of available features may be regarded as a search for an ideal feature mask. This is a binary mask, the same length as the complete feature vector, in which a 1 indicates a feature that is used and a 0 indicates a feature that is unused. The search space may therefore be expressed as a feature mask lattice [15], which is formed by successively flipping each bit of the feature mask.

In order to search this space, a means of assessing each feature mask and a method of performing the search are required. Classifier performance is often assessed through the use of a receiver operating characteristic (ROC) curve, which depicts true-positive rate against false-positive rate. However, a problem with this approach is that one cannot judge which of the two classifiers is superior if their ROC curves cross, because the relative superiority of one or the other depends on the misclassification cost that can be tolerated (and hence, the false-positive rate). To address this issue, the Parcel algorithm [27], which is a feature-selection algorithm for classification problems with a variable misclassification cost, is used. Parcel performs a heuristic search of feature mask space using the area under the ROC curve (AUROC) as a performance metric.

B. Parcel Algorithm

Parcel [27] allows for both bottom-up and top-down searching of the feature mask, depending on the initial state of the search. For example, for a bottom-up search of a three-attribute feature vector, the initial mask would be

$$\mathrm{POOL}_{\mathrm{init}} = \{[0,0,0]\}. \tag{4}$$

The successors of the initial feature mask are then determined by successively flipping a bit, leading to an initial mask pool expressed as

$$\mathrm{POOL}_1 = \{[1,0,0],[0,1,0],[0,0,1]\}. \tag{5}$$

The classification performance is then computed using the feature vectors corresponding to each feature mask in the pool and assessed using ROC curves. Each point on an ROC curve represents the performance obtained by setting a threshold on a

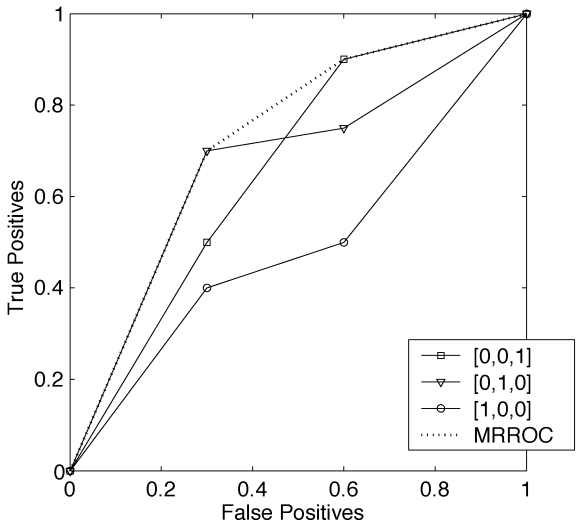

Fig. 7. Example of an MRROC derived from the ROC curves for three classifiers with different feature masks.

continuous output from the classifier (i.e., a classification confidence score). Hence, an ROC curve is a concave function generated by varying this threshold to produce different tradeoffs between true-positive and false-positive rates [31]. An example is shown in Fig. 7.

Furthermore, it is possible to interpolate between classifiers in the ROC space. Consider two classifiers, one that consistently reports a positive (thus having 100% true and false positives) and one that consistently reports a negative (thus having 0% true and false positives). These classifiers occupy points at (1,1) and (0,0) in the ROC space, respectively. Now, consider a new classifier that randomly selects one of these classifiers with an equal probability of selection; such a classifier would score 50% true and false positives. Similarly, by choosing classifiers with different probabilities, any point on the line in the ROC space between the two classifiers can be attained [27].

The notion of interpolating between two points in the ROC space can be extended to the example in Fig. 7, and hence, it is the convex hull over the points in the ROC space that describes the optimal performance. The convex hull is shown as a dashed line in Fig. 7 and is denoted as the maximum realizable ROC (MRROC). The MRROC indicates that optimal performance can be obtained by selecting points from the ROC curves of a number of different classifiers (i.e., the MRROC is a concatenation of ROCs).

During its search of feature mask space, Parcel only examines classifiers that lie on the MRROC. Therefore, in the example described above, the mask pool for the second iteration consists of

$$\mathrm{POOL}_2 = \{[1,1,0],[1,0,1],[0,1,1]\} \tag{6}$$

which is the set obtained by flipping bits in the two feature masks that contribute to the MRROC (i.e., [0,0,1] and [0,1,0]).

The search continues in this manner until there are no more feature masks to examine, or until there is no significant increase in the area under the MRROC. Hence, Parcel performs a heuristic search of the feature lattice that will find a local maximum in the feature space. The resulting set of classifiers may not be optimal, however, since the local maximum found by Parcel may be inferior to a mask that was not examined. Despite this limitation, Parcel was found to be an effective means for searching the feature space that is highly suited to variable cost environments.

VI. Classification Algorithm

The framework outlined above is complete except for a description of the specific classification algorithm. Here, the choice of algorithm was largely dictated by the small amount of training data available—in some cases as few as 20 exemplars of a class were available. In such situations, a discriminative classifier is preferred; here, an adaptation of the KNN algorithm is used [2].

In the KNN algorithm, class membership of a target vector is determined from the labels of the training vectors surrounding the target vector in the feature space. Specifically, to determine the label of the target vector, the distance (generally Euclidean) from the target vector to each element of the training set is computed. The k closest training vectors are chosen, and the target vector is then labeled as the mode of this subset of the training vectors. In cases where two or more classes occur with the same frequency in this subset, the label is randomly chosen from the competing classes.

Recall that in order to produce an ROC curve, a classifier that produces a continuously valued confidence score is required. Keller *et al.* [14] describe an extension to the KNN algorithm that allows a confidence measure to be obtained from the classifier. This is computed as a fuzzy membership value

$$\mu_i(\mathrm{Te}) = \frac{\sum_{j=1}^{K} u_{ij}\left(\frac{1}{\|\mathrm{Te}-\mathrm{Tr}_j\|^{\frac{2}{(\Omega-1)}}}\right)}{\sum_{k=1}^{K}\left(\frac{1}{\|\mathrm{Te}-\mathrm{Tr}_j\|^{\frac{2}{(\Omega-1)}}}\right)} \tag{7}$$

where $\mu_{\mathrm{i}}(\mathrm{Te})$ is the fuzzy membership of the test vector Te to class i, $\|\mathrm{Te}-\mathrm{Tr}_j\|$ is the Euclidean distance between the test vector and the training vector, K is the number of nearest neighbors required, and Ω defines the "fuzziness" of the function. u_{ij} defines the membership in the ith class of the jth training vector and is determined *a priori* from the labeled training data by

$$u_{ij} = \begin{cases} 0.51 + \left(\frac{N_i}{K_t}\right) \times 0.49 & m = i \\ \left(\frac{N_i}{K_t}\right) \times 0.49 & m \neq i \end{cases} \tag{8}$$

where K_t is the number of nearest neighbors used to compute the membership value, N_i is the number of neighbors found, which belong to the ith class, and m is the class of the jth labeled training vector Tr_j. Hence, two extensions are made to the standard KNN algorithm. Firstly, (8) indicates that each training vector is given a measure according to how well it represents its class. A training vector that is surrounded by members of its own class is given a high value of u_{ij}, whereas one that is isolated from other members of its class is given a low value of u_{ij}. Secondly, (7) makes use of a fuzzy parameter Ω, which determines the extent to which the distance between training and

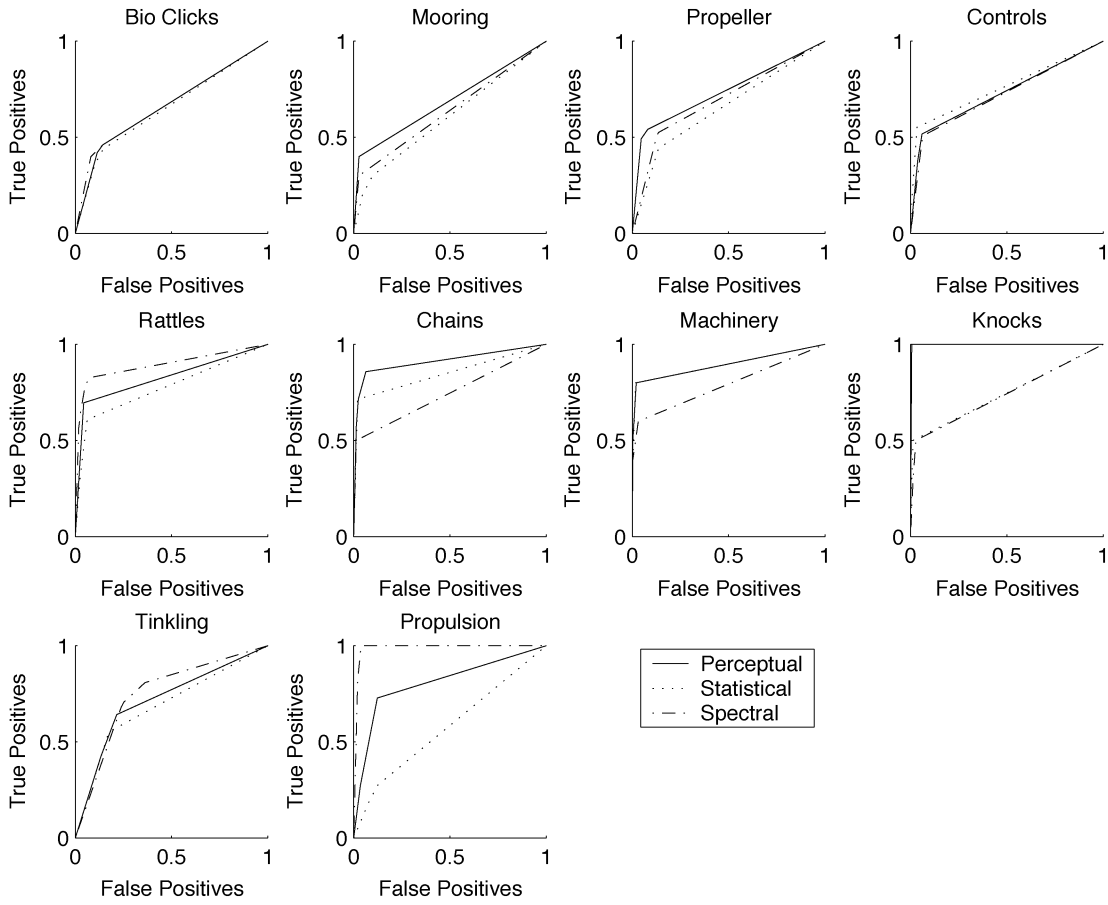

Fig. 8. Classification performance when Parcel was able to choose from spectral, statistical, or perceptual features only. Each panel shows an ROC curve for a transient class.

test vectors affects the confidence score. A value of $\Omega = 2$ was chosen on the basis of experiments with a small validation set.

VII. Evaluation

The classification architecture described above was evaluated on the corpus described in Section III using the following configurations: spectral features only, statistical features described by [26], perceptually motivated features described here, and all of these features in combination.

Spectral features were computed by filtering the acoustic signal with a bank of 16 gammatone filters [5], [7]. Filter center frequencies were spaced between 20 and 11 050 Hz on an ERB-rate scale. The mean energy in each channel was computed over the duration of the transient, giving a single vector of spectral features.

The features proposed by Ridge [26] are predominantly statistics in the first, second, third, and fourth order of time, frequency, and power. Measures of the rate of attack and decay are also included. Details are given in the Appendix.

The perceptually motivated features constituted the cues to timbre, material, and temporal pattern described above. Recall that the material estimation algorithm has a free parameter, which determines the width of the Gaussian used for smoothing when identifying spectral peaks. Rather than tune this parameter to a specific value, a range of ten smoothing values was used. The resulting ten sets of material features were included in the feature vector, so that Parcel could select features with an appropriate smoothing parameter for each class.

In the current study, only the problem of transient classification is addressed; it is assumed that transients are detected and endpointed by another process (for example, see [25]). Accordingly, the start and end points of transient events were identified manually, and short and long contexts were isolated. The short context (which delimited the transient *per se*) was used to compute the spectral, statistical, timbre, and material features. The long context (10 s) was used to compute the rhythmic features.

VIII. Results

Classifier performance was evaluated for a range of false-positive values, and thus results are presented in the form of ROC curves. The AUROC is used as a metric for summarizing the performance of a classifier. In general, better classifiers will have a larger AUROC (note, however, that for a given false-positive rate, the classifier with the largest AUROC may not have the highest true-positive rate). Here, we are primarily interested in comparing classifiers that use different acoustic features; to do so, the ratio of their AUROCs is reported. In the following, a ratio of AUROCs is calculated as X/Y, so a ratio greater than 1 indicates that X has a greater AUROC than Y.

A. Comparison of Feature Vectors

ROC curves for classifiers trained on the spectral, statistical, and perceptual features are shown in Fig. 8, and results for classifiers trained on all features are shown in Fig. 9. The results are summarized in Table II, in the form of a comparison of AUROC

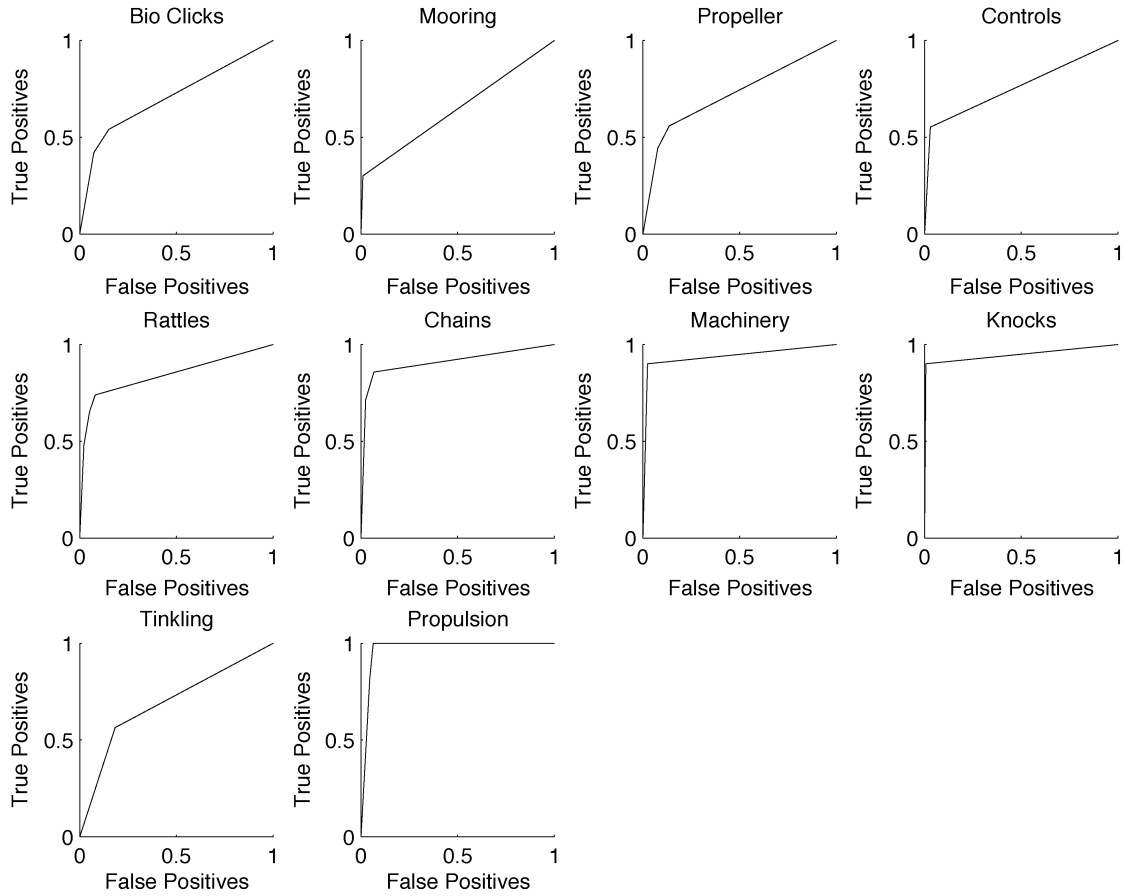

Fig. 9. Classification performance when Parcel was able to select from all available features. Each panel shows an ROC curve for a transient class.

TABLE II
CLASSIFIER PERFORMANCE FOR DIFFERENT FEATURE SETS EXPRESSED AS A RATIO OF AUROCS (X/Y). LARGER NUMBERS INDICATE BETTER PERFORMANCE ACROSS ALL CLASSES OF TRANSIENT

		X			
		Proposed	Statistical	Spectral	All
Y	Proposed	1.00	0.94	0.97	1.01
	Statistical	1.07	1.00	1.04	1.08
	Spectral	1.04	0.96	1.00	1.04
	All	0.99	0.93	0.96	1.00
	Totals	4.10	3.83	3.97	4.13

ratios. The results show that for the classification framework used here, the proposed perceptually motivated features are superior to both the statistical and spectral features. Furthermore, simple spectral features outperform the statistical features on this corpus, and the best performance is achieved when all of the features are made available for feature selection.

Examination of the ROC curves suggests that performance on the Bio Clicks class is relatively poor in all cases. It should be noted that this class was the most heterogeneous of those considered here, since it contained clicks emitted by a number of biological sources. Propulsion sounds are best recognized using spectral features, presumably because they have a characteristic spectral pattern that is not modeled well by the statistical or perceptual features.

Signals in the Knocks class have a characteristic temporal pattern (recall Fig. 5), which is apparently well represented by the proposed rhythmic features. For this class, the proposed features give a substantial improvement in performance compared to the spectral and statistical features.

In overall terms, the best classification performance is obtained when the Parcel algorithm is able to select from all of the available features. Note, however, that better performance was obtained on the Knocks class using the perceptually motivated features only. This highlights the fact that Parcel aims to optimize performance for all classes, rather than for each individual class. In the case of the Knocks class, Parcel failed to find the best performing feature set, which most likely consisted of perceptually motivated features only.

B. Analysis of Feature Selection

It is instructive to examine the selections made by Parcel when all features were available. Each ROC curve in Fig. 9 represents the performance of a number of classifiers, which may employ different features at different operating points. For a meaningful analysis, the feature set used at a particular operating point must therefore be chosen. Here, the operating point at which the ROC curve crosses the line connecting (1,0) and (0,1) is chosen, since points on this line have equal error rates (i.e., $\text{true positive rate} = 1 - \text{false positive rate}$). For each

TABLE III
FEATURES SELECTED BY PARCEL FOR EACH CLASS OF TRANSIENT. MA_n INDICATES THE nTH ELEMENT OF THE MATERIAL COMPONENT VECTOR, ST_n IS THE nTH STATISTICAL FEATURE (SEE THE APPENDIX) AND SP_n CORRESPONDS TO THE nTH SPECTRAL FEATURE. TE_{XX} IS THE LIKELIHOOD THAT CLASS XX WILL PRODUCE THE OBSERVED RHYTHMOGRAM FEATURES. SF AND LF REFER TO THE TIMBRAL CUES OF SPECTRAL FLUX AND LOWEST FREQUENCY PEAK, RESPECTIVELY. THE TEMPORAL CENTROID FEATURE WAS NOT USED BY ANY CLASS

Class	Selected features
Bio Click	$\mathrm{MA}_8\ \mathrm{ST}_7\ \mathrm{SP}_1\ \mathrm{SP}_3\ \mathrm{SP}_5\ \mathrm{SP}_{13}\ \mathrm{SP}_{16}$
Mooring	$\mathrm{SP}_1\ \mathrm{SP}_4\ \mathrm{SP}_5\ \mathrm{SP}_9\ \mathrm{SP}_{14}\ \mathrm{SP}_{16}$
Propeller	$\mathrm{ST}_3\ \mathrm{ST}_{10}\ \mathrm{ST}_{14}\ \mathrm{SP}_3\ \mathrm{SP}_5\ \mathrm{SP}_9\ \mathrm{SP}_{16}\ \mathrm{TE}_{\mathrm{BC}}\ \mathrm{TE}_{\mathrm{MO}}\ \mathrm{TE}_{\mathrm{RA}}\ \mathrm{TE}_{\mathrm{PN}}$ SF
Controls	$\mathrm{SP}_4\ \mathrm{SP}_7\ \mathrm{SP}_{10}$
Rattles	$\mathrm{ST}_4\ \mathrm{ST}_{23}\ \mathrm{SP}_5\ \mathrm{TE}_{\mathrm{KN}}\ \mathrm{TE}_{\mathrm{PN}}$ LF
Chains	$\mathrm{ST}_3\ \mathrm{ST}_9\ \mathrm{ST}_{21}\ \mathrm{ST}_{22}$ LF
Machinery	$\mathrm{MA}_1\ \mathrm{MA}_7\ \mathrm{ST}_3\ \mathrm{TE}_{\mathrm{KN}}\ \mathrm{TE}_{\mathrm{PN}}$
Knocks	$\mathrm{MA}_5\ \mathrm{ST}_5\ \mathrm{TE}_{\mathrm{CO}}\ \mathrm{TE}_{\mathrm{CH}}\ \mathrm{TE}_{\mathrm{MA}}\ \mathrm{TE}_{\mathrm{KN}}$
Tinkling	$\mathrm{ST}_1\ \mathrm{ST}_6\ \mathrm{ST}_7\ \mathrm{ST}_8\ \mathrm{ST}_{11}\ \mathrm{ST}_{14}\ \mathrm{ST}_{16}\ \mathrm{ST}_{21}\ \mathrm{ST}_{23}$
Propulsion	$\mathrm{MA}_5\ \mathrm{ST}_{23}\ \mathrm{SP}_1\ \mathrm{SP}_{14}\ \mathrm{SP}_{15}$

class, Table III shows the features used by the classifier that was closest to this operating point on the ROC curve.

Parcel selected between three and twelve features to represent each class of transient, with a mean of six features selected. Features were used from the perceptual, spectral, and statistical categories in almost equal proportion (20:22:22, respectively). Seven classes used perceptual features, although some (such as Controls and Tinkling) used statistical or spectral features alone. It is promising to note that classifiers which performed particularly well (e.g., those for the Rattles and Machinery classes) mainly used perceptual features. It is noted, however, that for this corpus, one of the features relating to timbre—temporal centroid—was not selected for any class.

The perceptually motivated features that encode rhythmic information appear to have been used in a way that is intuitively reasonable. For example, the Knocks class has a characteristic temporal structure, and uses four such features.

It is also interesting to note that rhythmogram features associated with a particular class have been employed by other classes. For example, the Propeller class uses features that encode the likelihood that the Bio, Mooring, Rattles, and Propulsion classes have produced the observed temporal pattern. In this case, the temporal structure of the Propeller class is substantially different from these other classes, and hence, strong evidence for, say, the Bio class will mitigate against the Propeller class.

IX. SUMMARY AND DISCUSSION

This paper has described a framework for classifying transient sonar signals, together with perceptually motivated acoustic features that encode properties such as timbre, material, and temporal pattern. The proposed approach is enforced by two observations. Firstly, sonar transients vary widely in their spectral and temporal properties, and hence, different classes of transient may be most effectively modeled by different acoustic features. Accordingly, a feature-selection approach based on the Parcel algorithm is used [27], which determines the optimal features for modeling each class. Secondly, transient sonar sounds often occur in a characteristic temporal pattern. Hence, cues that encode the temporal distribution of transient events over a relatively long time window (10 s) are employed.

The particular classifier used was a fuzzy KNN algorithm that allows a confidence measure to be derived. However, the proposed framework is quite general, and any classifier that produces a continuous output could be used. Similarly, other acoustic features could be used in place of, or in addition to, the ones described here.

The results of the experiments provide some justification for using perceptually motivated acoustic features. The perceptual features were superior to both the statistical and spectral features when used alone, although better performance was obtained by allowing Parcel to select from all of the available features. The results confirm that the classification of some transient sounds can be improved by employing features that encode the temporal context in which they occur (specifically, the temporal pattern). Features that encode perceptual attributes such as timbre and material also appear to be useful, although they were selected less frequently by Parcel and were not chosen for some classes. In some cases (e.g., the Mooring class), spectral features were chosen in preference to timbral features, suggesting that the latter did not provide an effective encoding of the characteristic spectral properties of the class. However, timbral features proved effective in modeling other classes such as Rattles and Chains.

The training phase of the proposed approach is relatively time consuming, since it requires the evaluation of multiple alternative classifiers for each class (feature selection using Parcel required approximately 3 days of processing time on a 500-MHz Pentium III machine). However, once the class models have been derived, the testing phase only requires the computation of the feature vectors and confidence scores. In principle, close to real-time performance could be achieved during testing (although it is noted that the 10-s window required for computation of the rhythmogram features is a limitation in this regard). A further advantage is that the training set used by the fuzzy KNN algorithm can be updated dynamically, if it is assumed that new exemplars are adequately represented by the acoustic features selected for the class.

It should be noted that the Parcel algorithm does not scale to an arbitrary number of features without concern for the amount of training data available. Specifically, very large amounts of training data are required if the pool of candidate features is large, because of the "curse of dimensionality" [2]. In the present simulations, Parcel selected an average of six features to represent each class. Given the limited training data available to us, it is possible that some classifiers whose feature masks contained many 1s were not properly characterized, and this was the reason why they were not selected. This issue will be investigated in future work.

The proposed approach is well suited to dealing with the variable cost of transient misclassification. When the cost of a misclassification is low (e.g., during routine ocean monitoring), an

operating point that gives a high true-positive rate at the cost of many false positives may be specified. Similarly, when the cost of a misclassification is high (e.g., during combat), a low incidence of false positives may be paramount. Such decisions are not affected during the training of the classifier; they can be made dynamically during its use.

Finally, it is noted that Collier [8] recently reported a psychophysical study in which novices and trained sonar operators were asked to categorize sonar recordings as being either of man-made or biological origin. He found that novices could perform this task almost as well as trained sonar operators, and did so using similar strategies for most signals. This result lends support to the proposed approach, because it suggests that general acoustical properties (such as timbre and temporal pattern) are likely to be exploited by both novices and trained listeners. It is therefore possible that the present system could model the findings from Collier's study. This remains to be tested empirically. It also remains to be shown whether this system would scale to the larger corpus used by Collier, which contained 23 classes of transient as opposed to the 10 classes used here.

As well as a comparison against human performance, future work will consider a wider range of perceptually motivated features. The extent to which the proposed approach is able to discriminate the acoustic signatures of merchant shipping from other sounds will also be investigated.

APPENDIX

This appendix details the statistical features proposed by Ridge [26], which are denoted as ST_n in Table III. During preprocessing, the signal is split into T frames and the STFT of size $2F$ is computed for each frame. The signal is therefore represented by T frames of F spectral coefficients. In the following, $p_{t,f}$ denotes the power at time t and frequency f.

Duration: the number of time frames in the signal

$$\mathrm{ST}_1 = T. \tag{9}$$

Peak power: maximum of the total power over all time frames

$$\mathrm{ST}_2 = M = \max(p_t) \tag{10a}$$

$$p_t = \sum_f p_{t,f}. \tag{10b}$$

Average power: mean of the total power over all time frames

$$\mathrm{ST}_3 = \frac{\sum_t p_t}{T}. \tag{11}$$

Time of peak power: the index of the time frame at which the peak power occurs

$$\mathrm{ST}_4 = L = \arg\max_t(p_t). \tag{12}$$

Frequency of peak power: the index of the frequency bin in which the peak power occurs

$$\mathrm{ST}_5 = \arg\max_f(p_f) \tag{13a}$$

$$p_f = \sum_t p_{t,f}. \tag{13b}$$

Mean frequency: the weighted mean of the frequency of the event, where P denotes the total power

$$\mathrm{ST}_6 = \bar{f} = \frac{\sum_f f \cdot p_f}{P} \tag{14a}$$

$$P = \sum_f p_f. \tag{14b}$$

rms bandwidth: a measure of the frequency bandwidth, where $*$ indicates element-wise multiplication

$$\mathrm{ST}_7 = B = \sqrt{\frac{\sum_f f^2 * p_f}{P} - \bar{f}^{\,2}}. \tag{15}$$

Frequency skew: the mean of the frequency skew

$$\mathrm{ST}_8 = \sqrt{\frac{\sum_f (f-\bar{f})^3 * p_f}{B^3 P}}. \tag{16}$$

Frequency kurtosis: the mean of the frequency kurtosis

$$\mathrm{ST}_9 = \sqrt{\frac{\sum_f (f-\bar{f})^4 * p_f}{B^4 P}}. \tag{17}$$

Mean time: the mean time of the event, weighted by power

$$\mathrm{ST}_{10} = \bar{t} = \left(\frac{\sum_t p_t t}{PT}\right). \tag{18}$$

rms time: a measure of the temporal bandwidth

$$\mathrm{ST}_{11} = C = \sqrt{\frac{\sum_t t^2 * p_t}{P} - \bar{t}^{\,2}}. \tag{19}$$

Temporal skew: the mean of the temporal skew

$$\mathrm{ST}_{12} \sqrt{\frac{\sum_t (t-\bar{t}^{\,3}) * p_t}{C^3 P}}. \tag{20}$$

Temporal kurtosis: the mean of the temporal kurtosis

$$\mathrm{ST}_{13} = \sqrt{\frac{\sum_t (t-\bar{t})^4 * p_t}{C^4 P}}. \tag{21}$$

Power SD: standard deviation of the power

$$\mathrm{ST}_{14} = \sqrt{\frac{FT\sum_{t,f}(p_{t,f})^2}{P^2}} - 1. \tag{22}$$

Power SDT: standard deviation of the power in time

$$\mathrm{ST}_{15} = \frac{T\sum_{t} p_t^2}{P^2} - 1. \tag{23}$$

Power SDF: standard deviation of the power in frequency

$$\mathrm{ST}_{16} = \frac{F\sum_{f} p_f^2}{P^2} - 1. \tag{24}$$

Power skew: skew of the power

$$\mathrm{ST}_{17} = \frac{\frac{1}{FT}\sum_{t,f}(p_{t,f} - \bar{P})^3}{\bar{P}^3} \tag{25a}$$

$$\bar{P} = \frac{P}{FT}. \tag{25b}$$

Power skewT: skew of the power in time

$$\mathrm{ST}_{18} = \frac{\frac{1}{T}\sum_{t}\left(\frac{p_t}{F} - \bar{P}\right)^3}{\bar{P}^3}. \tag{26}$$

Power skewF: skew of the power in frequency

$$\mathrm{ST}_{19} = \frac{\frac{1}{F}\sum_{f}\left(\frac{p_f}{T} - \bar{P}\right)^3}{\bar{P}^3}. \tag{27}$$

Power kurtosis: kurtosis of the power

$$\mathrm{ST}_{20} = \frac{\frac{1}{FT}\sum_{t,f}(p_{t,f} - \bar{P})^4}{\bar{P}^4}. \tag{28}$$

Power kurtosisT: kurtosis of the power in time

$$\mathrm{ST}_{21} = \frac{\frac{1}{T}\sum_{t}\left(\frac{p_t}{F} - \bar{P}\right)^4}{\bar{P}^4}. \tag{29}$$

Power kurtosisF: kurtosis of the power in frequency

$$\mathrm{ST}_{22} = \frac{\frac{1}{F}\sum_{f}\left(\frac{p_f}{T} - \bar{P}\right)^4}{\bar{P}^4}. \tag{30}$$

Rate of attack: maximum rate of increase of the total power in each time frame, from the start of the signal to its peak level

$$\mathrm{ST}_{23} = \max\left(\frac{p_t - p_{t-1}}{M}\right), \quad t = 0, \ldots, L. \tag{31}$$

Rate of decay: minimum rate of decrease of the total power in each time frame, from the peak level of the signal to its end.

$$\mathrm{ST}_{24} = \min\left(\frac{p_t - p_{t+1}}{M}\right), \quad t = L, \ldots, T. \tag{32}$$

ACKNOWLEDGMENT

Thanks to Peter Glynn and Ruth Wilcox of QinetiQ for their support and encouragement, and for their help in obtaining the corpus of sonar signals.

REFERENCES

[1] ANSI, Psychoacoustical Terminology S3.20, Amer. Nat. Standards Inst., New York, 1973.
[2] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford Univ. Press, 1995.
[3] B. Boashash and P. O'Shea, "A methodology for detection and classification of some underwater acoustic signals using time–frequency analysis techniques," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, pp. 1829–1841, Nov. 1990.
[4] S. F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Washington, DC, 1979, pp. 200–203.
[5] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, no. 4, pp. 297–336, 1994.
[6] J. D. Carroll and J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart–Young' decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.
[7] M. Cooke, *Modelling Auditory Processing and Organisation*. Cambridge, U.K.: Cambridge Univ. Press, 1993.
[8] G. L. Collier, "A comparison of novices and experts in the identification of sonar signals," *Speech Commun.*, vol. 43, no. 4, pp. 297–310, 2004.
[9] K. van den Doel and D. K. Pai, "The sounds of physical shapes," *Presence*, vol. 7, no. 4, pp. 382–395, 1998.
[10] D. J. Freed, "Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events," *J. Acoust. Soc. Amer.*, vol. 87, no. 1, pp. 311–322, 1990.
[11] B. L. Giordano, "Material categorization and hardness scaling in real and synthetic impact sounds," in *The Sounding Object*, D. Rocchesso and F. Fontana, Eds., 2003, pp. 73–94.
[12] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *J. Acoust. Soc. Amer.*, vol. 61, no. 5, pp. 1270–1277, 1977.
[13] S. Handel, "Timbre perception and auditory object identification," in *Hearing*, B. C. J. Moore, Ed. San Diego, CA: Academic, 1995, pp. 425–461.
[14] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE Trans. Syst., Man, Cybern.*, vol. 15, pp. 580–585, Jul. 1985.
[15] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1–2, pp. 273–324, 1997.
[16] A. Kundu, G. C. Chen, and C. E. Persons, "Transient sonar signal classification using hidden Markov model and neural nets," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Adelaide, Australia, 1994, pp. 325–328.
[17] A. J. Kunkler-Peck and M. T. Turvey, "Hearing shape," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 26, no. 1, pp. 279–294, 2000.
[18] T. S. Leung and P. R. White, "A fuzzy logic based underwater acoustic transient classifier," in *7th IEEE Digital Signal Proc. Workshop*, Loen, Norway, 1996, pp. 494–497.
[19] X. Li, R. J. Logan, and R. E. Pastore, "Perception of acoustic source characteristics: Walking sounds," *J. Acoust. Soc. Amer.*, vol. 90, no. 6, pp. 3036–3049, 1991.
[20] S. L. Marple, *Digital Spectral Analysis: With Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
[21] S. McAdams, "Perspectives on the contribution of timbre to musical structure," *Comput. Music J.*, vol. 23, no. 3, pp. 85–102, 1999.
[22] P. M. Oliveira, V. Lobo, V. Barroso, and F. Moura-Pires, "Detection and classification of underwater transients with data driven methods based on time–frequency distributions and non-parametric classifiers," in *Oceans 2002 MTS/IEEE*, vol. 1, Biloxi, MS, pp. 12–16.
[23] T. W. Parks and B. A. Weisburn, "Classification of whale and ice sounds with a cochlear model," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, San Francisco, CA, 1992, pp. 481–484.
[24] G. Peeters, S. McAdams, and P. Herrera, "Instrument sound description in the context of MPEG-7," in *Proc. Int. Computer Music Conf. (ICMC)*, Berlin, Germany, 2000.
[25] L. A. Pflug, G. E. Ioup, and J. W. Ioup, "Multichannel moment detectors for transients in shallow-water noise," *IEEE J. Ocean. Eng.*, vol. 29, no. 1, pp. 157–168, Jan. 2004.

[26] A. Ridge, "Scalar Parameter Extraction From FFT Data for Transient Association," Sound Concepts Department, QinetiQ, Winfrith, U.K., 2003.
[27] M. J. J. Scott, M. Niranjan, and R. W. Prager, "Parcel: Feature Subset Selection in Variable Cost Domains," Cambridge Univ. Eng. Dept., Cambridge, U.K., Tech. Rep. TR 323, 1998.
[28] A. Teolis and S. Shamma, "Classification of Transient Signals via Auditory Representations," Syst. Res. Center, University of Maryland, College Park, Tech. Rep. TR91-99, 1991.
[29] N. P. M. Todd, "The auditory 'primal sketch': A multiscale model of rhythmic grouping," *J. New Music Res.*, vol. 23, no. 1, pp. 25–70, 1994.
[30] N. P. M. Todd and G. J. Brown, "Visualization of rhythm, time and metre," *Artif. Intell. Rev.*, vol. 10, no. 3–4, pp. 253–273, 1996.
[31] H. L. Van Trees, *Detection, Estimation, and Modulation Theory: Part 1*. New York: Wiley, 1968.
[32] S. Tucker, "An ecological approach to the classification of transient underwater acoustic events: Perceptual experiments and auditory models," Ph.D. Dissertation, Dept. of Comput. Sci., Univ. of Sheffield, Sheffield, U.K., 2003.
[33] S. Tucker and G. J. Brown, "Modelling the auditory perception of size, shape and material: Applications to the classification of transient sonar sounds," in *Proc. 114th Conv. Audio Engineering Society*, Amsterdam, Netherlands, 2003.

Simon Tucker received the M.Eng. degree in software engineering, in 2000 and the Ph.D. degree in computer science, in 2003, both from the University of Sheffield, Sheffield, U.K.

Since 2004, he has worked at the Department of Information Studies, University of Sheffield, as a Research Associate on the EU Augmented Multi-Party Interaction (AMI) project. He has research interests in automatic recognition and browsing of audio, machine learning, and temporal compression of speech.

Guy J. Brown received the B.Sc. degree in applied science from Sheffield Hallam University, U.K., in 1988, the Ph.D. degree in computer science and the M.Ed. degree from the University of Sheffield, Sheffield, U.K., in 1992 and 1997, respectively.

He has been a visiting Research Scientist at LIMSI-CNRS (Paris), ATR (Kyoto), The Ohio State University and Helsinki University of Technology. He is currently a Senior Lecturer in computer science with the University of Sheffield. He has a long-established interest in computational auditory modelling, and also has research interests in automatic speech recognition and music technology. He has authored and coauthored more than 80 papers in books, journals and conference proceedings.