# Using an Aural Classifier to Discriminate Cetacean Vocalizations

by

Carolyn Binder

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
March 2012

DALHOUSIE UNIVERSITY

DEPARTMENT OF

PHYSICS AND ATMOSPHERIC SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Using an Aural Classifier to Discriminate Cetacean Vocalizations" by Carolyn Binder in partial fulfilment of the requirements for the degree of Master of Science.

Dated: March 26, 2012

Supervisor: _____

_____

Readers: _____

_____

DALHOUSIE UNIVERSITY

DATE: March 26, 2012

AUTHOR: Carolyn Binder

TITLE: Using an Aural Classifier to Discriminate Cetacean Vocalizations

DEPARTMENT OR SCHOOL: Department of Physics and Atmospheric Science

DEGREE: M.Sc. CONVOCATION: May YEAR: 2012



_____

Signature of Author

*Dedicated to my grandfather, Murray F. Ward – an advocate of my university education.*

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABSTRACT

To positively identify marine mammals using passive acoustics, large volumes of data are often collected that need to be processed by a trained analyst. To reduce acoustic analyst workload, an automatic detector can be implemented that produces many detections, which feed into an automatic classifier to significantly reduce the number of false detections. This requires the development of a robust classifier capable of performing inter-species classification as well as discriminating cetacean vocalizations from anthropogenic noise sources. A prototype aural classifier was developed at Defence Research and Development Canada that uses perceptual signal features which model the features employed by the human auditory system. The dataset included anthropogenic passive transients and vocalizations from five cetacean species: bowhead, humpback, North Atlantic right, minke and sperm whales. Discriminant analysis was implemented to replace principal component analysis; the projection obtained using discriminant analysis improved between-species discrimination during multiclass cetacean classification, compared to principal component analysis. The aural classifier was able to successfully identify the vocalizing cetacean species. The area under the receiver operating characteristic curve (*AUC*) is used to quantify the two-class classifier performance and the *M*-measure is used when there are three or more classes; the maximum possible value of both *AUC* and *M* is 1.00 – which is indicative of an ideal classifier model. Accurate classification results were obtained for multiclass classification of all species in the dataset ($M = 0.99$), and the challenging bowhead/ humpback ($AUC = 0.97$) and sperm whale click/anthropogenic transient ($AUC = 1.00$) two-class classifications.

# LIST OF ABBREVIATIONS AND SYMBOLS USED

| | |
|---|---|
| $\mathbf{A}$ | Transformation matrix used for principal component analysis |
| $\mathbf{x}'$ | Vector describing the location of a point in the transformed space |
| $\mathbf{a}$ | Vector describing a principal component |
| ADAC | Acoustic Data Analysis Centre |
| *AUC* | Area under the ROC curve |
| BBD | Bin-to-bin difference |
| $c$ | Number of classes |
| $C_{end}(n)$ | Criterion function used to find the end of a vocalization in the Kliewer-Mertins vocalization isolation algorithm |
| CFAV | Canadian Forces Auxiliary Vessel |
| $C_{start}(n)$ | Criterion function used to find the start of a vocalization in the Kliewer-Mertins vocalization isolation algorithm |
| DA | Discriminant Analysis |
| dB | Decibel |

| | |
|---|---|
| DRDC | Defence Research and Development Canada |
| $E_L$ | The estimated energy to the left of a timeseries sample in the Kliewer-Mertins vocalization isolation algorithm |
| $E_R$ | The estimated energy to the right of a timeseries sample in the Kliewer-Mertins vocalization isolation algorithm |
| ERB | Equivalent rectangular bandwidth |
| $E_{thresh}$ | Basilar membrane excitation due to background noise in the head |
| $E_{vocal}$ | Basilar membrane excitation due to the vocalization |
| $FP$ | False positive rate |
| $\mathbf{H}$ | Transformation matrix used for discriminant analysis |
| KM | Kliewer-Mertins |
| $L$ | Length of rectangular sliding window used in the Kliewer-Mertins vocalization isolation algorithm |
| $M$ | $M$-measure: multiclass performance metric |
| $\mathbf{m}$ | Sample mean vector |
| $MR$ | Misclassification rate |
| $n$ | Negative class |
| $N$ | Classifier assigned a vocalization to the negative class |
| $N'(ERB)$ | Perceptual loudness, measured in sones/ERB |
| $p$ | Positive class |

| | |
|---|---|
| $P(\mathbf{x}'\|C)$ | Likelihood probability that a vocalization belonging to class $C$ is located at position $\mathbf{x}'$ |
| $P(C\|\mathbf{x}')$ | Posterior probability: likelihood that a vocalization located at position $\mathbf{x}'$ belongs to class $C$ |
| PAL | Passive aural listening |
| PAM | Passive acoustic monitoring |
| PCA | Principal component analysis |
| PDF | Probability density function |
| $p_k$ | Relative variance in the first $k$ principal components or discriminant functions |
| PMSBR | Psychoacoustic maxima-to-spectral-bins ratio |
| $R$ | Threshold used to quantify risks associated with misclassification |
| ROC | Receiver operating characteristic |
| S & T | Science and Technology |
| SARA | Species At Risk Act |
| $\mathbf{S}_B$ | Between-class scatter matrix |
| $\mathbf{S}_B$ | Within-class scatter matrix |
| SBDS | Sub-band decay slope |
| $s_D$ | Discriminant score: a measure of how well separated class means are relative to their within-class variance |
| SoX | Linux Sound Exchange application |

| | |
|---|---|
| SPL | Sound pressure level |
| $T_c$ | Time constant for band-limited energy detection |
| $TP$ | True positive rate |
| $\mathbf{w}$ | Discriminant function |
| $W_N$ | Length of noise window for band-limited energy detection |
| $W_S$ | Length of signal window for band-limited energy detection |
| $\mathbf{x}$ | Vector describing the location of a point in the feature space |
| $\mathbf{X}$ | Matrix formed from eigenvectors of the covariance matrix |
| $\mathbf{y}$ | Vector containing normalized feature values for each vocalization in the training subset |
| $Y$ | Classifier assigned a vocalization to the positive class |
| $y_N$ | Estimate of noise level for band-limited energy detection |
| $y_s$ | Estimate of signal level for band-limited energy detection |
| $\alpha$ | Averaging coefficient for band-limited energy detection |
| $\Delta T$ | Time resolution of band-limited energy detection |
| $\lambda$ | Eigenvalue |
| $\pi(C)$ | Prior probability for class $C$ |
| $\Sigma$ | Sample covariance matrix |
| $\sigma^2$ | Sample variance |

# ACKNOWLEDGEMENTS

I want to first thank my supervisor, Paul Hines for the opportunity to work on this project. He acted as a mentor in numerous ways and provided me with many, and varied, pieces of wisdom. I have learned a lot from him. I would also like to acknowledge the co-supervisory role of Richard Dunlap – your experience as a graduate supervisor was a great asset. Thanks also to Chris Purcell for serving on my supervisory committee and as a thesis reader. Harm Rotermund, thank you for providing your insights as a thesis reader. I greatly appreciate the advice and friendship of my co-workers at DRDC. Stefan Murphy was an excellent source of aural classifier wisdom and someone who I could easily bounce ideas back-and-forth with. I often turned to Sean Pecknold for random bits of information. Discussions with both Stefan and Sean were very helpful and invaluable. I would like to acknowledge the acoustic analysis effort supplied by Akoostix Inc; Joe Hood and Ben Bougher greatly helped me to develop an understanding of Akoostix's software suite that was used for detecting marine mammal vocalizations.

I am deeply indebted to all the members of my family for their support and encouragement throughout my Masters degree. Mom and Dad have always advocated the benefits of a higher education and most importantly critical thought – so, thank you for guiding me. Markus, you have been an incredible lifeline throughout my university career, providing me with the grounding I required. Thank you for always being there and understanding and supporting my academic goals.

Over the course of my M.Sc. degree I received financial support from NSERC, DRDC – Atlantic, and Dalhousie University.

# CHAPTER 1 INTRODUCTION

There has been increasing concern about the impact of anthropogenic effects on marine mammals. Much of this concern is related to the use of active sonar [1], [2], [3], marine oil and gas exploration [2], [4], [5], and high volumes of shipping traffic [6], [7]. Reliable knowledge of marine mammal presence may help avoid the negative impacts due to the aforementioned activities. For example, the Cornell Bioacoustics Research Program [6] has implemented a network of hydrophones (i.e. underwater microphones) to monitor for the presence of critically endangered North Atlantic right whales in the heavily used shipping lanes leading to Boston Harbour. If a whale is detected, an alert informs ship captains in the region about the location of the whale and to reduce ship speed to minimize the risk of striking the whale. This type of warning system requires accurate identification of a specific marine mammal species and a reliable system capable of ignoring false detections.

Automatic detection of the presence of marine mammals has become increasingly important to the Royal Canadian Navy as it strives to maintain its responsibility to environmental stewardship. A science and technology (S&T) challenge identified in the Defence S&T Strategy [8] is the need to minimize military impact on the environment both in operational and training contexts. The goal is to "identify technologies that can help to protect the environment while minimizing detrimental effects on operational effectiveness." If marine mammals are known to be in the area of a Navy exercise, the Navy may take mitigating actions to avoid any potential harm to the animals. The presence of marine mammals also may cause a negative impact on military exercises by

imposing limitations on active sonar use due to environmental laws and regulations. Additionally, marine mammal vocalizations may generate false alarms on the passive sonar systems used by the military to monitor the operational environment. A robust system capable of reliably detecting the presence of marine mammals automatically, as well as identifying a large number of species, is required to limit the negative effects to and by marine mammals, without producing further demands on sonar operators.

A computer-based aural classifier was developed at Defence Research and Development Canada (DRDC) to classify active sonar echoes from target and clutter objects. Young and Hines [9] employed timbre-defining perceptual features to take advantage of anecdotal and experimental evidence that sonar operators can hear the difference between these two types of sonar returns. The automatic aural classifier was able to successfully distinguish between impulsive source returns from man-made metallic objects (targets) and naturally occurring clutter objects with an 82.3% probability of detecting the targets and only a 1.9% probability of false alarm. The features used by the classifier were based on a model for human auditory perception. By its nature, the human auditory system is a passive system; thus, it seems a logical step to apply the aural classifier to a passive sonar case. Marine mammal vocalizations provide a readily available source of passive sonar sounds with which to test the aural classifier.

The marine mammal dataset used for this research is composed of cetacean vocalizations. Cetaceans are the group of marine mammals belonging to the order *Cetacea*, which includes the sub-orders *Mysticete* (baleen whales) and *Odontocete* (toothed whales, dolphins and porpoises). Generally, baleen whales are physically large and produce relatively low frequency sounds, whereas odontocetes tend to be smaller than baleen whales and produce sounds of a higher frequency [10].

Traditionally, cetacean presence was detected by visual surveys; however, the ability to visually detect cetaceans is limited because visual observers can only see them when the animals are at the surface, during daylight hours and when weather does not greatly reduce visibility. Visual surveys are also limited in both spatial and temporal scales [11]

because they require a dedicated and time-consuming effort, typically requiring the use of a ship. The use of passive acoustics (i.e. recording of sounds from the environment) has many advantages over visual survey methods – an acoustic recorder may be left in place for extended periods of time and continue recording regardless of time of day or weather conditions.

Passive acoustic monitoring (PAM) is now widely used for long-term survey efforts that investigate seasonal movements and identify critical habitat areas, or onboard ships as part of avoidance and sound mitigation systems [12]. Although PAM has the potential to allow for 24-hour, real-time, automated monitoring in all types of conditions, it does possess its own set of limitations. Marine mammals must vocalize to be detected (typically not a problem because most cetaceans are highly vocal) and those vocalizations must be distinguishable from the background noise and other sources, [13], [14]. Passive acoustic surveys also generate large volumes of data; for example a five-year survey for Shell Offshore Incorporated in Alaska's Beaufort and Chukchi Seas generated approximately 5 TB of acoustic data – to review such a large amount of data would require an expert acoustic analyst an estimated five "person years" of effort [4]. This example further highlights the need for automatic detection and classification systems to reduce the human analyst workload involved in analyzing acoustic data from PAM efforts.

Successful PAM depends on several factors including: sufficiently high vocalization rates by the species of interest; appropriate detection ranges – which in turn depend on source levels of vocalizations and environment dependant propagation characteristics; automatic detection/classification algorithms that are robust to noise and are able to identify variable signals; and a good understanding of the limitations of a particular detection/classification algorithm in terms of error rates [15]. The goal of PAM is to maximize the number of true whale detections and at the same time to minimize the number of missed detections and false positives [16].

Accurate classification of species at risk is required to determine population size and identify critical habitats to aid in developing mitigation strategies. As Arctic ice melts new shipping lanes may be developed through the Arctic Ocean. Knowledge of critical habitats as well as real-time detection of species may help to protect Arctic marine mammal populations that are identified as being sensitive due to warming trends and habitat loss [17]. Endangered marine mammal species are not exclusive to the Arctic, but also exist in other ocean basins including the high-use shipping lanes off the Eastern coast of North America.

In this thesis, aural classification is employed to address the issue of reliable automatic classification of marine mammals. Vocalizations from five cetacean species and a selection of anthropogenic transients were used to quantify the aural classifier's performance. Following the introduction, CHAPTER 2 reviews the theory associated with detection, aural classification and the metrics used to assess performance. CHAPTER 3 discusses the cetacean dataset in detail. Pre-processing of the data and detection of vocalizations is discussed in CHAPTER 4. In CHAPTER 5 results of inter-species aural classification of cetacean vocalizations are presented. CHAPTER 6 examines the classification of sperm whale clicks and anthropogenic transients. Implementation of discriminant analysis and comparison with principal component analysis results are discussed in CHAPTER 7. Some final conclusions are drawn in CHAPTER 8 in addition to suggestions for future research.

# CHAPTER 2 THEORY

## 2.1 AURAL CLASSIFICATION

The aural classifier developed by Young and Hines [9] was motivated by anecdotal evidence that experienced sonar operators could *hear* differences in active sonar returns from man-made metallic objects (targets) and naturally-occurring geologic objects (clutter). The novel aspect of the aural classifier is the type of signal features employed – features intended to provide cues for aural discrimination. The classifier architecture that is employed takes advantage of Gaussian-based statistics. This type of architecture is simple and is widely used in the pattern classification community because Gaussian statistics are easy to define and understand conceptually. The simplicity of the classifier architecture allows analysis to focus on the effectiveness of the aural features.

The perceptual signal features used by the aural classifier are derived from musical acoustics research – where much effort has been applied to identify signal characteristics that define timbre. "Timbre is that attribute of auditory sensation in terms of which a subject can judge that sounds similarly presented and having the same loudness and pitch are dissimilar [18]." For example, the perceived difference in sound between the musical note middle C being played on a violin and cello, with the same loudness and duration, is said to result from a difference in timbre.

Allen *et al.* [19] compared performance results of the aural classifier with results from human listening tests. This experiment also provided the opportunity to experimentally validate humans' ability to aurally discriminate active sonar returns. Human listeners

were presented with active sonar returns from target and clutter objects and were asked to identify each return based on how it sounded. Listeners were also asked to rank the certainty of each classification decision. The accuracy of the human listeners and the degree of certainty in making a classification decision were compared to the accuracy of the aural classifier and the likelihood probabilities associated with the classifier's decision. The comparison of human listening results to automatic classifier results was performed using only individual perceptual features. Some of the features ranked the returns in a similar way as the human listeners, which provided preliminary evidence that the perceptual features employed by the aural classifier correlate with features used by the human auditory system.

As outlined in the flow diagram in Figure 2.1, aural classification consists of several steps: the start and end of each vocalization is identified, a relatively simple audio model is applied and perceptual features are extracted from each vocalization, then the data is split into training and testing subsets; from the training set the most important aural features are identified and then used to perform classification using a Gaussian classifier. The classifier model determined from the training set is then applied to the testing set. Further details about each of these steps are provided in Sections 2.1.1 - 2.1.5. Young

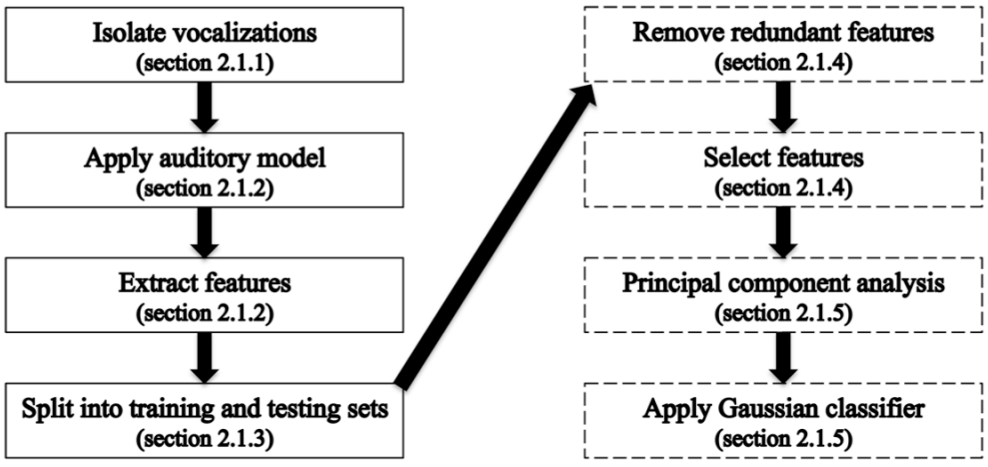


Figure 2.1 Diagram showing the steps of the classification process. Steps that are in dashed blocks are computed from the training subset – the results of these steps are then applied to the testing subset. The text in parentheses refers to the section in which each step will be further discussed.

and Hines [9], and Young [20] outlined the aural classification process for the two-class (binary) classification problem. The aural classifier has been generalized to perform classification of two or more classes in the current research. In the following sections, where there are differences in implementation between the binary and multiclass (i.e. three or more classes) aural classifiers, the binary case will be presented first. This is because, in general, the binary classification case is simpler both mathematically and conceptually.

### 2.1.1 Vocalization Isolation

The first step of aural classification is to isolate each vocalization from the ambient noise. Vocalizations are first detected in their original recordings that range from a couple of minutes to approximately 30 minutes, as described in Section 4.1. After this preliminary detection step, each vocalization is placed in the centre of a WAV file with noise context before and after the vocalization. A single WAV file exists for each detected vocalization.

The start and end of each vocalization, relative to the beginning of its WAV file, was identified using the Kliewer-Mertins [21] (KM) technique. The KM technique was originally developed for use in audio sub-band coding schemes to extract a transient signal using energy estimation. The KM technique is based on the idea that the transition from noise to signal is defined by a rapid change in energy level. To quantify this change in signal level, two rectangular sliding windows of length $L$ are used to estimate the signal level to the left ($E_L$) and right ($E_R$) of each sample, $n$, in a WAV file as follows,

$$E_L(n) = \frac{1}{L}\sum_{k=n-L}^{n-1} x^2(k) \quad \text{, and} \qquad \text{Eqn. 2.1}$$

$$E_R(n) = \frac{1}{L}\sum_{k=n+1}^{n+L} x^2(k) \quad , \qquad \text{Eqn. 2.2}$$

where $x(k)$ is the WAV file amplitude[1] at sample $k$. It was found that $L = 1024$ worked well for isolating the transients used in this research. The criterion functions used to define the start, $C_{start}(n)$, and end, $C_{end}(n)$, of a vocalization are computed as,

$$C_{start}(n) = \log\left(\frac{E_R(n)}{E_L(n)}\right)E_R(n) \quad \text{, and} \qquad \text{Eqn. 2.3}$$

$$C_{end}(n) = \log\left(\frac{E_L(n)}{E_R(n)}\right)E_L(n) \quad . \qquad \text{Eqn. 2.4}$$

The samples at which the maximum of these criterion functions occur are defined as the start and end of the signal, respectively [9].

The KM technique works well for short impulsive transients, for which this technique was designed; however, it does not work as well for longer amplitude modulated baleen whale vocalizations, since the transition from noise to signal does not result in as marked a change in energy as is characteristic of impulsive sounds like sperm whale clicks. Thus, the original KM technique was used for isolation of sperm whale clicks and anthropogenic transients; a modified version of the technique was developed for the baleen whale vocalizations. Due to large variations in energy during a baleen whale vocalization, the maximum values for $C_{start}$ and $C_{end}$ do not necessarily correspond to the start and end of the call. Instead, the original KM technique may erroneously select some small portion of the call as shown in Figure 2.2. To overcome this, a modified version of the KM technique was implemented for all baleen whale vocalizations.

The modified KM technique computes the values of $C_{start}$ and $C_{end}$ as given by Eqn. 2.3 and Eqn. 2.4, but instead of selecting the maximum value, these values were compared to a pre-defined threshold value (through experimentation, the unitless value 0.02, was found to work well for baleen vocalizations). The search for the endpoints of the

---

[1] Note that $E_L(n)$ and $E_R(n)$ are not true energy values (i.e. measured in Joules), but are a measure of the variance of the signal, and thus are *proportional* to energy. Since the purpose of the Kliewer-Mertins technique is to quantify the relative change in signal level, an exact energy value is not required. Additionally, the link between physical pressures and amplitude values of the WAV file timeseries is somewhat arbitrary due to the nature of the scaling implemented by the auditory model in succeeding steps. Therefore, no physical units for $E_L(n)$, $E_R(n)$, $C_{start}(n)$, and $C_{end}(n)$ are reported in this thesis.

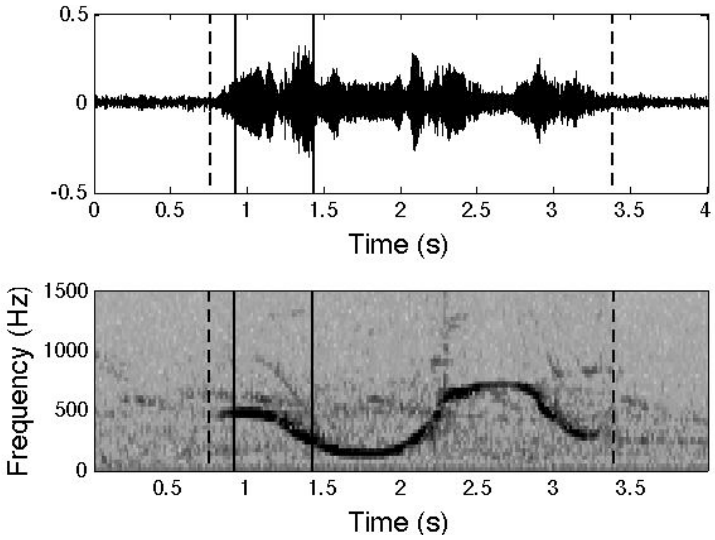


Figure 2.2 Timeseries and spectrogram of a bowhead vocalization with start and end points selected by the original Kliewer-Mertins technique (solid lines) and the modified version (dashed lines). The modified version selected the entire vocalization, whereas the original version selected approximately 20% of the vocalization.

vocalization begins in the centre of the WAV file, since it is known that the vocalization was placed approximately in the centre of the file. The search then moves outwards, away from the centre, until the criterion values are less than the threshold value. The samples at which $C_{start}$ and $C_{end}$ become less than the threshold value are defined as the beginning and ending of the vocalization. When results of the modified and original version of the KM technique are compared, as in Figure 2.2, it is clear that the modified version does a significantly better job of finding the true start and end samples of baleen whale vocalizations – the modified version captured the whole duration, whereas the original implementation captured only 20% of the vocalization. An accurate selection of the start and end of each vocalization is required because the time difference between these points (i.e. duration of vocalization) is used as a feature for classification.

### 2.1.2 Auditory Model and Feature Extraction

An auditory model that provides a perceptual representation of each vocalization is required to obtain the perceptual features used for classification. The auditory model employed by the aural classifier is relatively simple and is intended to produce a

perceptual model of each vocalization without being too computationally complex. The auditory model consists of three steps: apply the auditory filter bank, obtain the basilar membrane excitation pattern, and compute the perceptual loudness spectrum.

Since the auditory model provides a perceptual representation of each vocalization, many of the calculations are performed using perceptual units that do not necessarily directly correspond to standard physical units. For example, the sone unit is used to measure loudness; 1 sone has the same perceived loudness as a 1 kHz tone presented at 40 dB SPL, and as such, the definition of a sone is frequency dependent. A sound of *n* sones is perceived to be *n* times louder than a one sone sound. Due to the conversion to a perceptual representation of sound, some unitless scaling factors are required in Eqn. 2.5 and Eqn. 2.6.

Psychophysical models of human hearing assume the inner ear can be modelled as a bank of bandpass filters that are used to process sound [18]. The auditory filter bank employed is based on Slaney's [22] implementation of the Patterson-Holdsworth auditory filter bank. The filter bank is composed of 100 parallel bandpass filters (or channels) with centre frequencies between 20 and 4000 Hz, equally spaced on the equivalent rectangular bandwidth (ERB) scale. The ERB value defines the width of a filter channel (dependent on the centre frequency); in terms of linear frequency, the filter channels centred at lower frequencies will be narrower than filters centred at higher frequencies [23]. The ERB scale is representative of the manner in which humans perceive frequency and is defined in terms of, *f*, the conventional linear frequency expressed in units of Hz, as

$$ERBS = 21.4\log(0.00437f + 1). \qquad \text{Eqn. 2.5}$$

The value of *ERBS* indicates the number of ERBs (i.e. channels of the filter bank) below a particular frequency [23], *f*, required to provide appropriate spacing between adjacent channels of the auditory filter bank. The gain of each filter bank channel is scaled to represent propagation of sound through the outer and middle ear, and for complete audibility. Scaling for complete audibility ensures the perceptual signal spectrum of the signal exceeds the human hearing threshold throughout the frequency band of interest. For this research, the frequency band of interest ranges from the 20 Hz [18] low-

frequency limit of human hearing to the 4000 Hz Nyquist rate. Scaling each vocalization for complete audibility ensures that the extracted perceptual features will be based only on features audible to a human listener [9]. Time-frequency perceptual signal features – sub-band attack, sub-band decay, and sub-band synchronicity (see Appendix A for a definition of these features) – are obtained from the temporal envelopes of each scaled filter bank channel.

The basilar membrane converts the mechanical vibrations due to sound pressure into neural responses, which can be processed by the brain. The width and stiffness of the basilar membrane vary along its length, so the point of maximum vibration due to a sound stimulus varies depending on the frequency of the sound [18]. Thus, the basilar membrane performs as a frequency selector. In the auditory model employed by the aural classifier, the basilar membrane is modelled as 100 discrete points that correspond to the output of the auditory filter bank. A steady-state (i.e. time-invariant) model of the basilar membrane pattern is obtained by integrating the energy in each sub-band of the auditory filter bank. Since time dependence has been removed during this step, a purely spectral representation of the signal is obtained.

The final step of the auditory model is to apply a non-linear compression to the basilar membrane excitation pattern to obtain the perceptual loudness spectrum. The conversion to a perceptual loudness spectrum was proposed by Moore and Glasberg and can be expressed as

$$N'(ERB) = C\left[E_{vocal}(ERB)^{\alpha} - E_{thresh}(ERB)^{\alpha}\right], \qquad \text{Eqn. 2.6}$$

where $N'$ is the perceptual loudness in sones/ERB, $E_{vocal}$ represents the basilar membrane excitation due to the vocalization, $E_{thresh}$ is the basilar membrane excitation caused by background noise in the head (e.g. blood flow), α defines the non-linear compression (based on empirical results α = 0.2159), and $C$ is a scaling factor with a value of 0.0702 (obtained using trial-and-error by Young [20]) to ensure consistency with the definition of the sones scale. Since each vocalization was scaled for complete audibility, $E_{vocal} > E_{thresh}$ for all vocalizations and all ERB values. It is from the resulting perceptual loudness spectrum that purely spectral signal features – peak loudness, loudness centroid,

and loudness roughness – are extracted (see Appendix A for description of how features are extracted).

Each time-frequency feature calculated from the auditory filter bank can be thought of as a 100-dimensional feature, with the high dimensionality resulting from the 100 channels of the filter bank. Further treatment of these features for classification requires each feature to be one-dimensional; thus each time-frequency feature needs to be converted into representative one-dimensional features. This is accomplished through use of summary statistics. For each time-frequency feature the minimum, maximum and mean values are computed across all filter sub-bands. Additionally, the centre frequency of the filter bank channel corresponding to the minimum and maximum values are also treated as one-dimensional features. In this way, each 100-dimensional time-frequency feature is converted into five one-dimensional features. This treatment is unnecessary for the purely spectral features because they are already one-dimensional.

In summary, there are 46 time-frequency signal features that are extracted from the output of the auditory filter bank and 12 purely spectral signal features are derived from the perceptual loudness spectrum, giving a total of 58 one-dimensional perceptual signal features. See Appendix A for a full list of the perceptual signal features as well as a description of how each feature is computed.

### 2.1.3 Training and Testing Split

At this point in the process the dataset is split into two subsets – one is used to train and the other to test the classifier. The accepted method to determine a classifier model, based on the accumulated experience of many classifier developers and many different classification tasks, is to learn from example patterns. The aural classifier makes use of supervised learning [24] by providing a class label for each vocalization in the training set. By generalizing the underlying patterns in the training set, predictions can be made about data that do not have a known class label. Thus, the classifier model (discussed further in the following sections) is *trained* with vocalizations for which the classifier is

provided a class label. The effectiveness of the classifier is then *tested* by imposing the assumptions of the classifier on a dataset for which the classifier has no direct knowledge of the class label. To evaluate the performance of the classifier, the class label assigned by the classifier to each vocalization in the testing set can be compared to the known true label.

There is no duplication of vocalizations between the training and testing subsets, that is, any vocalization assigned to the training subset will not exist in the testing subset and vice versa. The training and testing split is accomplished by randomly selecting 50% of the vocalizations for each species to belong to the training subset and the remaining vocalizations for that species are placed in the testing subset. In this way, the training/testing split is performed species-by-species.

### 2.1.4 Feature Selection

Each feature is treated as an axis of the feature space. To ensure equal weighting of features and that final results are independent of feature value units, each feature is normalized to have a mean value of zero and variance of one. After normalization, each feature is unitless. Each vocalization can be treated as a point that is located within the feature space based on its feature values. The 58-element vector **x** describes the location of a sample vocalization in the feature space.

Some of the perceptual signal features may be equivalent to each other. There is a possibility that pairs of features provide the same information about the patterns that characterize each class of vocalization. These pairs of features are considered redundant – including only one of these redundant features is sufficient for quantifying the pattern. Removing the redundant features reduces computational complexity.

To identify redundant features, the sample correlation coefficient is computed between all feature pairs. The absolute values of the sample correlation coefficients are then compared to an established redundancy threshold (for this research a threshold of 0.9 is

used). All pairs of features whose correlation coefficients exceed the redundancy threshold contain at least one redundant feature. Features that occur most often in redundant pairs are removed from further consideration.

It is inevitable that some of the non-redundant features will be more useful for distinguishing between classes than others; for example, bowhead and humpback calls may have similar durations but very different peak loudness values. The training dataset can be used to form a subset of features that are the most useful for discriminating between classes.

There are many accepted methods for selecting the best features for classification – the method that is employed by the aural classifier is based on the Fisher Linear Discriminant method. The Fisher Linear Discriminant [24] is designed to select the features that best discriminate between classes. Suppose that the feature space is *d*-dimensional and $\mathbf{y}_1, \ldots, \mathbf{y}_d$ are the *d* *n*-dimensional sample vectors where *n* is the number of samples in the training set. Each vector **y** contains the normalized feature values for each vocalization in the training set. The Fisher discriminant score, $s_D$, is calculated for each feature as follows,

$$s_D = \frac{(m_1 - m_2)^2}{\sigma_1^2 + \sigma_2^2} \quad , \qquad \text{Eqn. 2.7}$$

where $m_1$ and $m_2$ are the sample means of the given feature separated by class (the subscript indicates to which class the mean value belongs) and $\sigma_1^2$ and $\sigma_2^2$ are the sample variances. For the multiclass case ($c > 2$ where $c$ is the number of classes), the Fisher score can be generalized to consider the separation between all class means relative to the dataset variance as follows,

$$s_D = \frac{\sum_{i=1}^{c} (m_i - m)^2}{\sum_{i=1}^{c} \sigma_i^2} \quad , \qquad \text{Eqn. 2.8}$$

where $m$ is the sample mean of the whole dataset, and $m_i$ and $\sigma_i^2$ are respectively the sample mean and sample variance of the $i^{th}$ class.

Features that do a good job of distinguishing between the classes will have sample class means that are well separated relative to the overall variance of the data and, therefore, will produce large $s_D$ values. Features with the largest values of $s_D$ are selected for inclusion in the feature subset used by the classifier. The number of features used for classification is determined on a case-by-case basis.

### 2.1.5 Classifier Architecture

A high-dimensional feature space requires many training samples to accurately estimate the underlying patterns in the data. Since the marine mammal dataset is limited, principal component analysis (PCA) is used to linearly combine the selected features to reduce the dimensionality. PCA is used on the training data to obtain a linear transformation that projects the feature space ($d$-dimensional) onto a new $k$-dimensional space, where $k \leq d$. $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the $n$ $d$-dimensional sample vectors with elements containing the normalized feature values [24]. The $d$ x $d$ sample covariance matrix, $\Sigma$, is computed as

$$\Sigma = \frac{1}{n}\mathbf{X}\mathbf{X}^T \quad , \qquad \text{Eqn. 2.9}$$

where the matrix $\mathbf{X}$ is formed by placing each of the $n$ vectors $\mathbf{x}$ into the rows of $\mathbf{X}$. Note that the bold notation represents variables that are vectors or matrices – this is standard throughout the thesis. The eigenvalues and eigenvectors of the sample covariance matrix are computed and sorted according to decreasing eigenvalue. The eigenvalue represents the relative amount of variance in the dataset captured by a particular eigenvector – eigenvectors corresponding to large eigenvalues lie in a direction, that when the data is projected onto the vector, maintain a large percentage of the dataset's variance. The relative amount of variance corresponding to $k$ of the $d$ total eigenvectors, or principal components, can be determined using the following relation,

$$p_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \quad , \qquad \text{Eqn. 2.10}$$

where $\lambda_i$ corresponds to the $i^{\text{th}}$ eigenvalue [20].

The transformation matrix, **A**, is formed from the $k$ eigenvectors corresponding to the $k$ largest eigenvalues, where $k$ is the desired number of PCA dimensions (typically $k = 2$). Thus, **A** will be a $d$ x $k$ matrix. To represent the data by the principal components, the data is projected onto the $k$-dimensional subspace according to the transformation,

$$\mathbf{x}' = \mathbf{A}^T \mathbf{x} \quad . \qquad \text{Eqn. 2.11}$$

Using Eqn. 2.11, principal component analysis restricts attention to the $k$ directions along which the scatter of the data points is greatest [24]. The axes of the resulting $k$-dimensional space are a linear combination of the selected features.

Unless otherwise noted, two principal components are used in this research. The selection of two principal components is somewhat arbitrary since it is possible to include any number of principal components up to and including the number of selected features. Two principal components are chosen primarily because two-dimensional spaces are easily represented graphically.

Within the PCA space, a Gaussian-based classifier is used. A Gaussian probability density function (PDF) is fit to each of the classes. For example, when classifying bowhead and humpback vocalizations, a Gaussian PDF is calculated for each whale species. A Gaussian classifier is used because it is recognized as the most common type of classifier and is the simplest to implement [24]. The Gaussian PDFs are represented as likelihood probabilities for a point $\mathbf{x}'$ in the PCA space, where $\mathbf{x}'$ is a two-element column vector. The likelihood that a bowhead vocalization would be located at position $\mathbf{x}'$ is given by,

$$P(\mathbf{x}'|B) = \frac{1}{2\pi|\Sigma_B|^{1/2}} \exp\left[-\tfrac{1}{2}(\mathbf{x}'-\mathbf{m}_B)^T \Sigma_B^{-1}(\mathbf{x}'-\mathbf{m}_B)\right] \quad , \qquad \text{Eqn. 2.12}$$

and the likelihood that a humpback vocalization would be located at position $\mathbf{x}'$ is expressed as,

$$P(\mathbf{x}'|H) = \frac{1}{2\pi|\Sigma_H|^{1/2}} \exp\left[-\tfrac{1}{2}(\mathbf{x}'-\mathbf{m}_H)^T \Sigma_H^{-1}(\mathbf{x}'-\mathbf{m}_H)\right] \quad . \qquad \text{Eqn. 2.13}$$

The variables $\Sigma_B$ and $\Sigma_H$ represent the 2 x 2 element sample covariance matrices for bowhead and humpback vocalizations in the training set. The sample mean vectors, $\mathbf{m}_B$

and $\mathbf{m}_H$ are the estimated mean values of the bowhead and humpback distributions, respectively.

Bayesian decision theory is used to combine the likelihood probabilities with prior probabilities, $\pi(B)$ and $\pi(H)$. These prior probabilities are calculated directly from the training set and represent the relative number of vocalizations produced by each species. Using Bayes' theorem, the posterior probability for bowheads is found as,

$$P(B|\mathbf{x}') = \frac{\pi(B)P(\mathbf{x}'|B)}{\pi(B)P(\mathbf{x}'|B) + \pi(H)P(\mathbf{x}'|H)} \quad , \qquad \text{Eqn. 2.14}$$

and the humpback posterior probability is calculated using,

$$P(H|\mathbf{x}') = \frac{\pi(H)P(\mathbf{x}'|H)}{\pi(B)P(\mathbf{x}'|B) + \pi(H)P(\mathbf{x}'|H)} \quad . \qquad \text{Eqn. 2.15}$$

These posterior probabilities are used to classify data in the testing set by substituting the location of each vocalization in the subset into Eqn. 2.14 and Eqn. 2.15. The ratio of the posterior probabilities is used to inform the classification decision. The decision rule dictates what decision to make for any given observation – for every value of $\mathbf{x}'$, the decision rule is able to assign either the bowhead or humpback class label [24]. The decision rule is based on the conditional risk, that is, the risk associated with misclassifying the whale species. If the risk associated with misclassifying a bowhead were much higher than for misclassifying a humpback then – in an effort to correctly identify the largest number of bowhead whales – the decision rule would require that $P(H|\mathbf{x}')$ be much greater than $P(B|\mathbf{x}')$ in order for a vocalization in the testing set to be classified as a humpback. Throughout this thesis, it has been assumed that there are equal risks associated with misclassifying any of the species of interest. Because of the assumption of equal risk, the specific decision rule used is to decide a vocalization to be from a bowhead if $P(B|\mathbf{x}') > P(H|\mathbf{x}')$, otherwise it is decided to be from a humpback. Figure 2.3 graphically illustrates how the Gaussian posterior probability distributions are used to form a 2D decision region; the decision boundary (i.e. the line between the red and blue regions) depicts the points of equal likelihood probability, $P(B|\mathbf{x}') = P(H|\mathbf{x}')$.

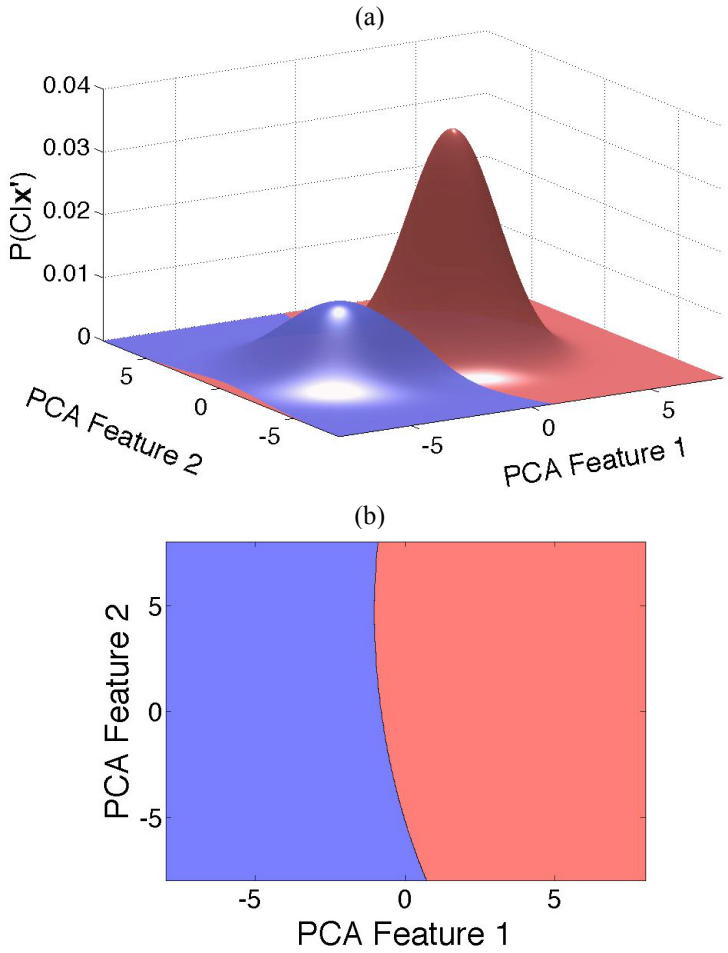


Figure 2.3 Illustration of how a decision region is generated using a Gaussian-based classifier. Bivariate Gaussian likelihood probability distributions, as in (a), are used to form the decision region shown in (b). The boundary between the red and blue areas of the decision region is defined by equal likelihood probabilities.

The multiclass case is simply a generalization of the binary case. For class *C* the posterior probability is calculated as,

$$P(C|\mathbf{x}') = \frac{\pi(C)P(\mathbf{x}'|C)}{\sum_C \pi(C)P(\mathbf{x}'|C)} \quad . \qquad \text{Eqn. 2.16}$$

The parallel with Eqn. 2.14 and Eqn. 2.15 is clear – there were only two possible classes in the bowhead and humpback binary example, so the denominator consisted of only two terms. In general, there will be as many terms in the denominator as number of classes. Since there are equal risks associated with each class, the decision rule is determined by the largest posterior probability.

### 2.1.6 Discriminant Analysis Theory

An alternative method to PCA, for projecting data onto a new *k*-dimensional space, is discriminant analysis (DA). DA seeks the combination of features that allows the best separation of class means while maintaining relatively little within-class variance. PCA is the dimensionality reduction method employed by the aural classifier throughout most of this thesis; however, CHAPTER 7 will examine the effects on classifier performance that result from implementation of DA. This section outlines the theory of DA and how it is implemented.

As was the case for PCA, the feature space is *d*-dimensional, and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the *n* *d*-dimensional sample vectors with elements containing the normalized feature values determined from data in the training subset. The vector $\mathbf{m}$ is the *d*-dimensional mean vector, corresponding to the mean of all the data points. As discussed in Section 2.1.4, features are normalized, based on the training subset, so that each feature has zero mean and variance of one. In the case of *c* classes (where $c \geq 2$) it is possible to compute $c - 1$ discriminant functions. The projection resulting from DA reduces the *d*-dimensional space to at most a $c - 1$ dimensional space, where it is assumed that $d \geq c$.

To obtain good separation of the projected data, the separation of the class means should be large relative to some measurement of the variance of each class. The between-class scatter matrix,

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \quad , \qquad \text{Eqn. 2.17}$$

is a measure of the separation between each class mean, $\mathbf{m}_i$, where $n_i$ is the number of samples in the $i^{\text{th}}$ class. Similarly, the within-class scatter matrix,

$$\mathbf{S}_W = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T \quad , \qquad \text{Eqn. 2.18}$$

provides a measure of the distance between each sample and its class mean. The optimal projection will result from a trade-off between minimizing within-class scatter and maximizing the separation of class means, which can be found by solving the following eigenequation,

$$\mathbf{S}_B\mathbf{w} = \lambda_i\mathbf{S}_W\mathbf{w}_i \quad . \qquad \text{Eqn. 2.19}$$

For this eigenequation to have a solution, $\mathbf{S}_W$ must be non-singular (i.e. the matrix must have an inverse). Because $\mathbf{S}_B$ is the sum of $c$ matrices of at most rank one, where only $c$ - 1 of these matrices are independent, then $\mathbf{S}_B$ will be at most of rank $c$ - 1. Thus, there will be no more than $c$ - 1 non-zero eigenvalues. The resulting set of eigenvectors, $\mathbf{w}_i$, corresponding to the non-zero eigenvalues, are the discriminant functions [24].

The discriminant functions form the columns of the transformation matrix, **H**, used to project the data onto a $c$ - 1 dimensional subspace as follows,

$$\mathbf{x}' = \mathbf{H}^T\mathbf{x} \quad . \qquad \text{Eqn. 2.20}$$

Similar to PCA, it is possible to select a subset of the $k$ best discriminant functions (i.e. the discriminant functions corresponding to the $k$ largest eigenvalues), for example, to facilitate graphically representing data in a two-dimensional subspace. The following ratio can be used to measure the degree of separation maintained after reducing the number of discriminant functions,

$$p_k = \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{c-1}\lambda_i} \quad . \qquad \text{Eqn. 2.21}$$

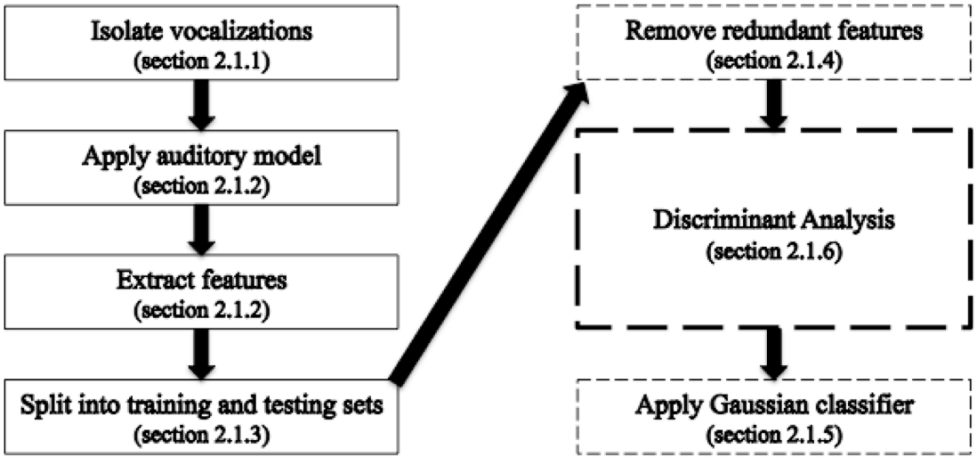


Figure 2.4 Steps of the classification process; steps that are in dashed blocks are computed from the training subset – the results of these steps are then applied to the testing subset. The bold dashed block, representing DA, replaces the Feature Selection and PCA blocks in Figure 2.1. The text in parentheses refers to the section in which each step is discussed.

Implementation of DA replaces the feature selection and PCA steps in the automatic classifier. Refer to the flow diagrams in Figure 2.1 and Figure 2.4 to see the differences in the classification process when DA replaces PCA.

## 2.2 PERFORMANCE METRICS

To evaluate classification effectiveness, appropriate performance metrics must be utilized. The simplest method for judging classifier performance is to qualitatively analyze the scatter and overlap of data points in the two-dimensional PCA space and determine how many misclassifications have been made versus how many correct classifications have been made. To visually verify that a correct classification has been made, one needs only to determine if a given sample has been plotted within the associated decision region. Figure 2.5 displays an example decision region for the simplest case – binary classification. Correct classification occurs when the crosses fall on the grey area and the circles fall within the white region. There are four incorrect classifications in this example, which are represented by the four circles that have been plotted on the grey decision region. Note that although it may seem that simply moving

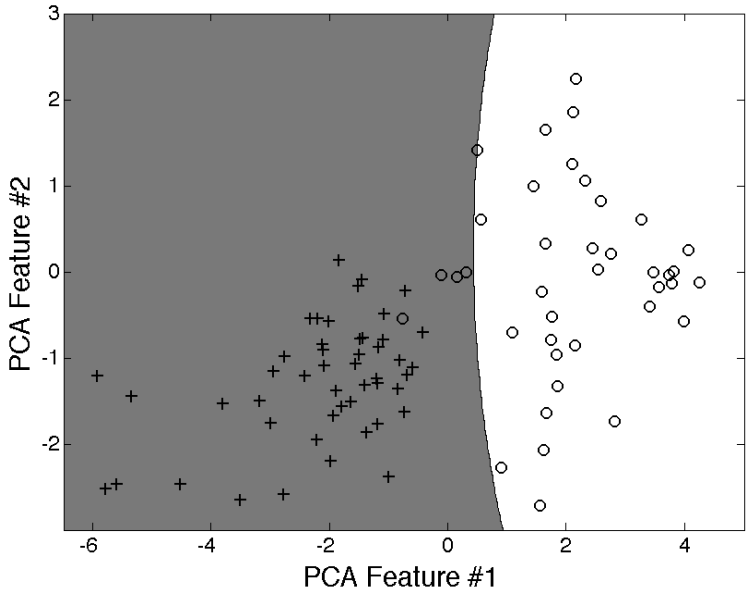


Figure 2.5 Example decision region displaying results from the test subset. Correct classification occurs when crosses are on the grey region and circles on the white region. Note that all samples have been correctly classified except for the four samples shown as circles that have been plotted on the grey region.

the decision boundary to the left would result in additional correctly classified points, it should be remembered that the decision boundary is determined from the training subset (refer to Figure 2.3 to view how the decision region is generated using Gaussian PDFs fit to data in the training subset). In this case, all data points in the training set were correctly classified. From this example plot it can be seen that plotting 2D decision regions provides a simple qualitative method to visually assess classifier performance; nonetheless, more sophisticated metrics are also required to provide a quantitative representation of classifier performance.

The simplest classification problem has only two true classes (a positive class, *p*, and a negative class, *n*) and two options for the classifier decision (assigned to the positive class, *Y*, or assigned to the negative class, *N*). There are four possible outcomes of a classification decision (summarized in, what is commonly referred to as, the "confusion matrix" shown in Figure 2.6): if the instance is positive and is classified as positive it is considered a "true positive"; if instead the positive instance is classified as negative it is counted as a "false negative". Alternatively, if a negative instance is classified as negative it is considered a "true negative", whereas if it is classified as a positive it is counted as a "false positive" (sometimes referred to as a "false alarm"). Figure 2.6 shows the confusion matrix of the possible outcomes for a particular instance and classification decision, as well as the common performance metrics that can be calculated [25].

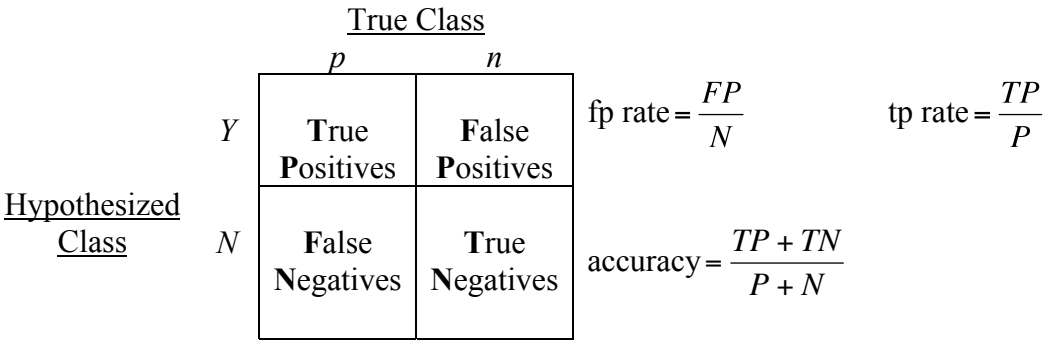


Figure 2.6 Confusion matrix for the two-class classification problem and the performance metrics that can be calculated. The symbols *p* and *n* represent the truth-value (positive or negative) of an instance, whereas *Y* and *N* represent the decision assigned by the classifier (after Ref. [25]).

Perhaps the most intuitive measures of classifier performance are the classifier accuracy and misclassification rate, where accuracy and misclassification rate (*MR*) are related by $MR = 1 - accuracy$; however, both these values are sensitive to class skew (i.e. different number of instances in the positive class and the negative class) and assume that the misclassification risks are equal [26], thus when either is used as a performance metric, classification should be restricted to the case of equal samples and risk associated with each class. Alternatively, the true positive and false positive rates can be combined at various levels of risk to generate a receiver operating characteristic (ROC) curve – this will be discussed in the next section.

### 2.2.1 ROC Curves

ROC curves illustrate the relative tradeoffs between the benefits (true positives) and costs (false positives) of a classifier model [25], by generating plots with false positive rate (*FP*) on the horizontal axis and true positive rate (*TP*) on the vertical axis (see Figure 2.7). The classifier computes the probability that an instance is a class member, which can be used to form the following decision rule, as discussed in Section 2.1.5,

$$\frac{P(n|\mathbf{x}')}{P(p|\mathbf{x}')} > R \quad , \qquad \text{Eqn. 2.22}$$

where the numerator is the probability the instance belongs to the negative class, the denominator is the probability the instance belongs to the positive class, and *R* is referred to as the threshold value. The value *R* is used to quantify the risks associated with misclassification.

For $R = 0$ all instances will be classified as negatives such that $P(p) = P(n) = 0$, producing the point on the ROC graph in the lower left corner. The point at the top right corner of the ROC curve is produced when $R = \max\left[P(n|\mathbf{x}')/P(p|\mathbf{x}')\right]$, so that all instances will be classified as positives and $P(p) = P(n) = 1$. By varying the threshold value between these two extremes and keeping track of the number of true positives and false positives at each *R*, a ROC curve is traced out [9].

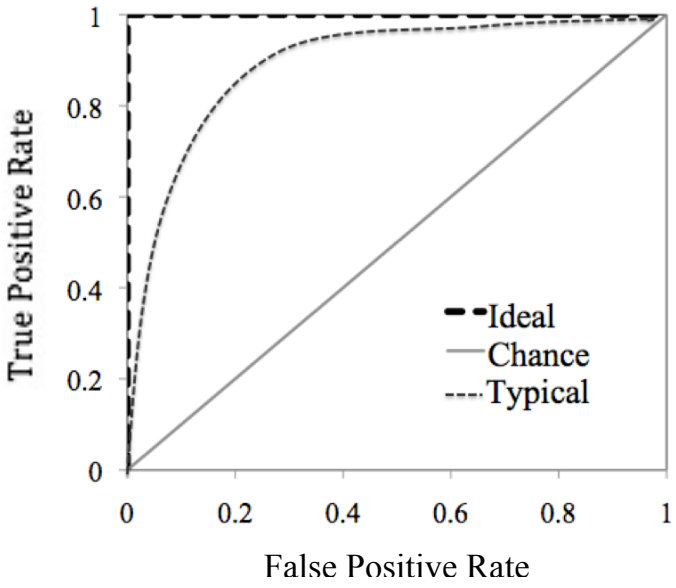


Figure 2.7 Examples of the ideal, chance, and what might be considered a typical ROC curve. Note that the area under the ideal curve is 1.0, and the area under the chance curve is 0.5.

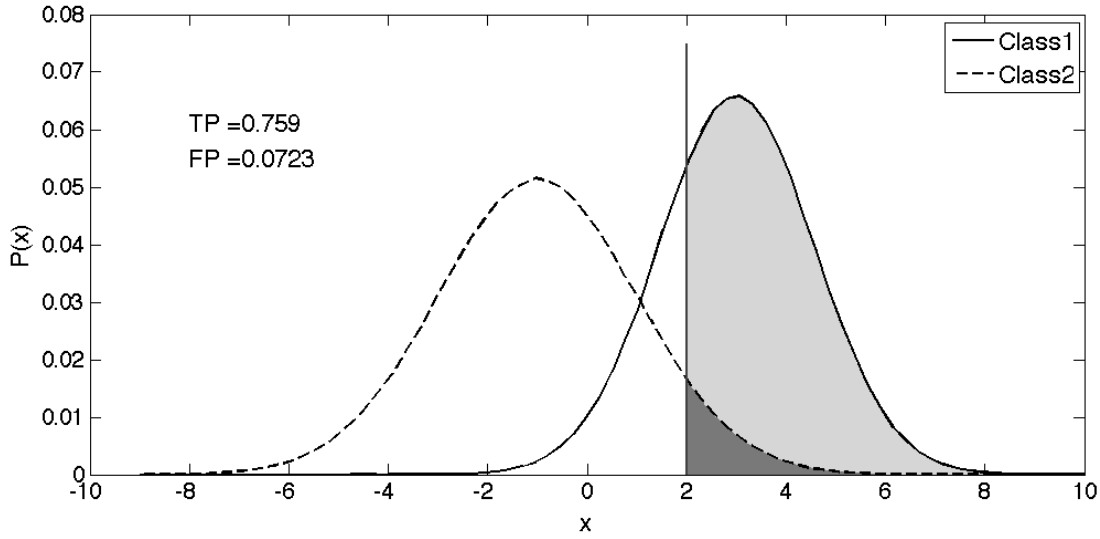


Figure 2.8 Example ROC curve generation from normal distributions – this example corresponds to a one-dimensional decision region. The solid, vertical line at $x = 2$ depicts the threshold value. In this case, Class1 represents the positive class and Class2 the negative class. The area under the Class1 curve, to the right of the threshold (all shaded areas), represents the true positive rate; the area under the Class2 curve, to the right of the threshold (dark shaded area), represents the false positive rate.

Continuing with the assumption of Gaussianity discussed in Section 2.1.5, a Gaussian PDF is fit to each of the vocalization classes. Figure 2.8 shows a snapshot of the ROC curve generation by depicting a single threshold value. This snapshot produces a single point on the ROC curve corresponding to the shaded areas under the PDFs to the right of the threshold. In the case shown, a single point would be plotted at (0.0725, 0.759). The complete ROC curve is obtained by smoothly varying the position of the threshold across the width of the Gaussians to encompass the points (0,0) and (1,1) on the ROC curve. A single value of $R$ is selected for a given classifier model – for example, if the risk associated with missing a true positive is large, then the threshold can be set with a relatively low threshold to flag any vocalization of interest [19] – this would correspond to a single point on the ROC curve.

Examples of ideal, chance and a typical ROC curve are shown in Figure 2.7. The straight line of the chance ROC curve is obtained by using a classifier that randomly guesses class membership. To move away from this line toward the ideal curve, a classifier must exploit information in the dataset. ROC curves provide the benefits of insensitivity to changes in class distribution, and visualization of classifier performance at various risk values [25], that are not possible by examining a simple decision surface.

The equal error rate is a relatively simple measure of classifier performance that can be estimated from the ROC curve and indicates when the risks associated with false positives and false negatives are equal. The equal error rate is defined as the point on the ROC curve at which $FP = 1 - TP$ (note that $FN = 1 - TP$, where $FN$ is the false negative rate); this is the point at which the ROC curve intersects with a diagonal line connecting the points (0, 1) and (1, 0). Smaller values for the equal error rate indicate better performance: an ideal classifier will have an equal error rate of 0%, whereas a classifier that randomly assigns class labels will have an equal error rate of 50%. In practice, the ROC curve is not continuous so the equal error rate is determined by finding the $FP$ that minimizes the value of $|1 - TP - FP|$.

### 2.2.2 Area Under ROC Curve

For quantitative comparison of classifiers, the information present in the ROC curve can be reduced to a single value – the area under the curve – *AUC* [25]. The *AUC* also provides a measure of how much the class distributions differ; large values indicate that the class distributions are substantially different [26]. The area under the ideal ROC curve is $AUC = 1$ and the chance ROC curve has $AUC = 0.5$. Typical ROC curves have *AUC*s ranging from the chance value to the ideal; no realistic classifier has $AUC < 0.5$, since the ROC curve can simply be inverted by reversing the positive and negative classification decisions. An important statistical property of the *AUC* is that it is equivalent to the probability that the classifier will rank a randomly selected positive instance higher than a randomly selected negative instance. In general, a larger *AUC* value implies better average classifier performance, although it is possible that a ROC curve with smaller *AUC* will perform better at some threshold values [25].

It is difficult to determine all possible sources of error in calculating the *AUC*; however, the error is likely dominated by the size of the dataset and the selection of the training subset. The *AUC* can be interpreted as the percentage correct in a forced-choice test [27], i.e. the probability of correctly classifying a randomly drawn sample from the dataset. Based on this interpretation of the *AUC* a limit on the number of significant figures can be placed according to the size of the test dataset. Table 3.3 contains the total number of vocalizations in the entire dataset. Due to the number of vocalizations per species (in the 100 – 500 range), no more than two significant figures should be used when presenting *AUC* results.

To quantify the error associated with choice of the training subset, a simple experiment was designed to estimate the variance in *AUC* results. Bowhead and humpback vocalizations were randomly selected and placed in the training subset and the remaining vocalizations placed in the testing set. This resulted in an approximately equal number of vocalizations per species in the training and testing subsets. Classification was performed as usual and the *AUC* value was noted. This process was repeated 100 times, each time randomly selecting vocalizations for the training and testing subsets. The

standard deviation of *AUC* values was 0.01. Since these two species have the largest variation in vocalization type, an upper-bound estimate on the standard deviation of *AUC* due to choice of training subset was set at 0.01. Therefore, all *AUC* values in this thesis will be presented with two significant figures to reflect the two largest sources of error in estimating the *AUC*.

### 2.2.3 Multiclass AUC

When more than two classes are considered a single ROC curve cannot be used to evaluate classifier performance. With more than two classes the resulting space is more complex to manage; for $c$ classes the confusion matrix analogous to that shown in Figure 2.6 becomes a $c$ x $c$ matrix with a total of $c$ correct classifications on the main diagonal and $c^2 - c$ possible errors (off-diagonal entries of the confusion matrix) [25]. The confusion matrix of *AUC* values for the $c$ class case is shown in Table 2.1, with classes labelled 1, 2, 3, …, $c$, ($c > 2$). There are a total of $c^2 - c$ entries in the confusion matrix corresponding to the number of possible errors. This is a symmetric matrix with $AUC(i, j) = AUC(j, i)$. Because $AUC(i, j) = AUC(j, i)$, there are a total of $\left(c^2 - c\right)/2$ unique errors. The entry $AUC(i, j)$ in the confusion matrix is the probability that a randomly selected instance belonging to class $j$ will have a lower estimated probability of belonging to class $i$ than a randomly selected member of class $i$ [26] – note that this is analogous to the definition of the binary *AUC* that was defined in section 2.2.2. Although the total proportion of correct classifications may be small, the decision rule employed

Table 2.1 Confusion matrix of areas under pairwise ROC curves for classification of $c$ classes, when $c > 2$. There are no entries on the main diagonal because it is not possible to classify a class against itself.

| | **Class 1** | **Class 2** | **Class 3** | **…** | **Class *c*** |
|---|---|---|---|---|---|
| **Class 1** | | *AUC(1,2)* | *AUC(1,3)* | *AUC(1,j)* | *AUC(1,c)* |
| **Class 2** | *AUC(2,1)* | | *AUC(2,3)* | *AUC(2,j)* | *AUC(2,c)* |
| **Class 3** | *AUC(3,1)* | *AUC(3,2)* | | *AUC(3,j)* | *AUC(3,c)* |
| **…** | *AUC(i,1)* | *AUC(i,2)* | *AUC(i,3)* | *AUC(i,j)* | *AUC(i,c)* |
| **Class *c*** | *AUC(c,1)* | *AUC(c,2)* | *AUC(c,3)* | *AUC(c,j)* | |

may be very accurate for certain classes or groups of classes [26], which can be seen by examining the entries in the confusion matrix.

The performance of the classifier in separating all $c$ classes – termed the $M$-measure – is given by,

$$M = \frac{2}{c(c-1)} \sum_{i<j} AUC(i,j) \quad . \qquad \text{Eqn. 2.23}$$

Thus, the $M$-measure is the average over all the pairwise $AUC$ values listed in the confusion matrix. The multiclass generalization of the $AUC$, as implemented by Hand and Till [26], has the same properties as the binary case – no information about class priors is required, classification is not limited to the case of equal costs associated with misclassification, and it yields a measure of how well each class is separated from all other classes. Like the $AUC$, the possible values for the $M$-measure range between 0.5 and 1.0, with larger values indicating superior classifier performance than lower values of $M$.

# CHAPTER 3 CETACEAN DATASET

Vocalizations from five cetacean species form the dataset used for aural classification. The frequency range of the vocalizations of each species significantly overlaps the human auditory range (20 Hz – 20 kHz for undamaged hearing [28]) so a clear parallel can be drawn between human expert listeners aurally classifying whale vocalizations and the results of the aural classifier. For comparison, the vocalization frequency ranges for the selected species are shown in Figure 3.1. In most cases, the frequency bandwidths of each species' vocalizations overlaps with the bandwidths of the other species, so frequency alone will not function well as a discrimination cue.

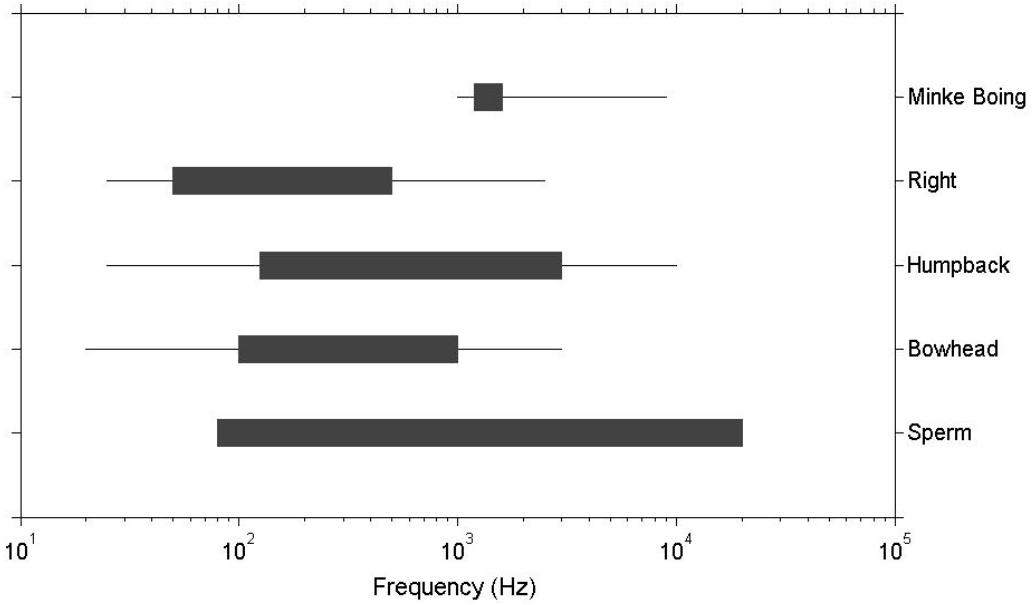


Figure 3.1 Known frequency ranges (plotted on a logarithmic scale) of cetacean vocalizations of the selected species. The thick bar shows the frequencies of the most common types of vocalizations and the thin line shows recorded frequency extremes [11], [29].

Five different species were selected for this research: bowhead, humpback, North Atlantic right, minke and sperm whales. An important factor in selecting these species was the availability of audio data containing many vocalizations. As discussed in the following sections, other factors in species selection included the endangered or threatened status of these species in the Canadian Species At Risk Act (SARA; see Table 3.1 for definition of these listings), particularly for the critically endangered North Atlantic right whales. The sperm whale was selected because automatic detectors often confuse its clicks with impulsive anthropogenic transients. Bowhead and humpback vocalizations are so similar in their frequency bandwidth and duration that many automatic classification algorithms have difficulty distinguishing between these sounds, and so were selected to provide a challenging case for the aural classifier.

Table 3.1 Definitions of the Canadian Species at Risk Act (SARA) categories used to describe the status of species in the Canadian wild [30]. The SARA categories are organized from highest risk to lowest risk to a wildlife species' survival.

| **SARA Category** | **Definition** |
|---|---|
| Extinct | A wildlife species that no longer exists. |
| Extirpated | A wildlife species that no longer exists in the Canadian wild but exists elsewhere in the wild. |
| Endangered | A wildlife species that is facing imminent extirpation or extinction. |
| Threatened | A wildlife species that is likely to become endangered if nothing is done to reverse the factors leading to declining population size. |
| Special Concern | A wildlife species that is likely to become threatened or endangered due to a combination of biological characteristics and identified threats. |

The dataset for this research was compiled using vocalizations that originated from various sources. Although there is no single accepted comprehensive catalogue of typical marine mammal vocalizations [31], some researchers have attempted to generate databases of vocalizations for use in automatic classification research – one such website is MobySound [32]. MobySound is a freely available reference archive that was constructed to facilitate research on automatic recognition of marine mammal sounds by providing a set of common vocalizations that researchers can use to test the effectiveness of various classifier implementations and compare results to those of other classification methods [33]. MobySound relies on researchers providing relatively good quality data

with accurate metadata associated with the recordings and vocalizations contained. MobySound is somewhat limited because it may not contain vocalizations from a particular species or a specific population of interest. For example, MobySound contains sounds from the North Pacific but not North Atlantic right whale; because cetacean vocalizations often differ between geographically separate populations of a species, it cannot be assumed that classification results from the North Pacific right whale can be directly applied to North Atlantic right whales. When possible, vocalizations from MobySound were used; otherwise vocalizations were obtained from other sources. The source of each species dataset is discussed in Sections 3.2 - 3.6.

Some of the species in the dataset produce a wide variety of sounds (e.g. humpbacks); in these cases a subset of vocalizations were selected. Aurally distinct vocalizations were included in the subset of a species' vocal repertoire only if a relatively large number of the sounds were available. Julie Oswald and Christine Erbe (both employed by JASCO Research [34]) were consulted to provide their expertise during selection of representative sounds.

When generating a dataset for automatic classification research, it is important to establish reliable ground truth information for the data to ensure that the classifier is trained with the correct data. At present, human experts remain the best classifiers, so experts in marine mammal vocalizations provided ground truth classifications using aural and visual (i.e. spectrogram-based) methods in addition to prior contextual information to confirm the species producing each vocalization used in this research [31].

## 3.1 WHALE SONG STRUCTURE AND TERMINOLOGY

The hierarchical organization of vocalizations used by some baleen whale species is referred to as whale song. Payne and McVay [35] wrote a defining paper describing the songs of humpback whales – this definition and terminology has been extended to the three other baleen whale species that have been reported to produce song. Of the singing baleen whales, humpbacks have been the most studied in large part due to the ease with

which researchers can do research in areas where humpback whales sing (e.g. near Hawaii and Puerto Rico) [10], so they are often used as examples when discussing the nature of whale song. This section will provide a brief overview of whale song intended to introduce the terminology that will be used in this thesis.

Whale song follows a hierarchical model. A "unit" is the shortest continuous sound perceived by human hearing that is preceded and succeeded by brief moments of silence. In some cases when a unit is played back at slower speed, or the fine-scale structure of the spectrogram is analyzed, it can be noted that the unit is actually made up of shorter segments of discrete pulses or tones, which can be referred to as "subunits." A series of units is called a "phrase", groups of similar phrases form a "theme," and a "song" is made up of several distinct themes. The term for a series of songs sung in succession, with less than one second between songs, is "song session." Song sessions can last for hours and a single song may last between 5 minutes and more than 30 minutes [35]. Figure 3.2

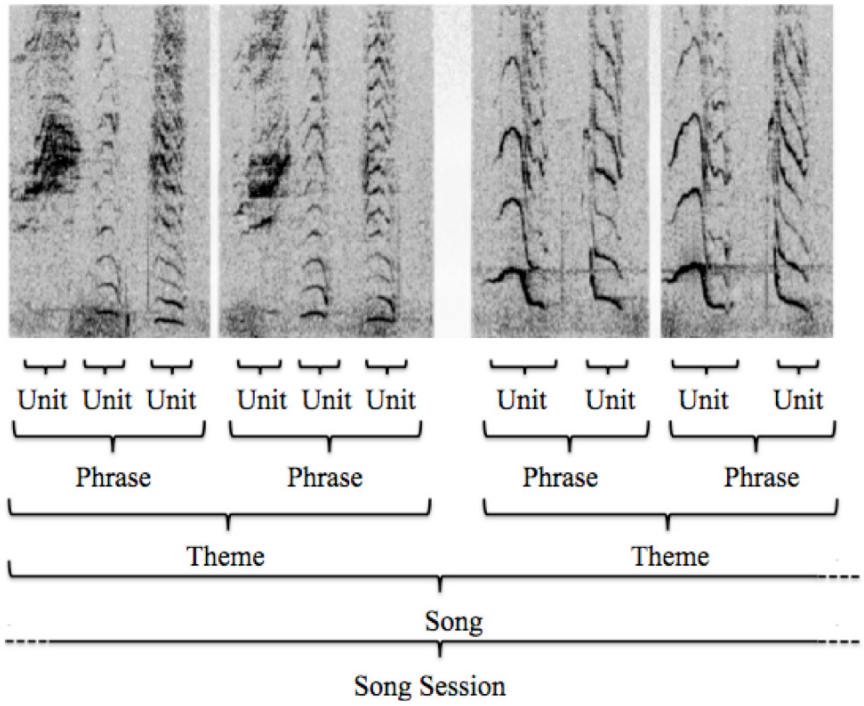


Figure 3.2 Diagram containing example spectrograms of humpback whale song and the associated terminology used to describe song. Frequency is on the vertical axis and time on the horizontal axis (after Ref. [35]).

shows a graphical representation of whale song organization with humpback whale spectrogram examples.

There is a high degree of repetition in a whale song. Only a few units are repeated in a certain order to form a phrase, and the phrase is repeated several times within a theme, and so on. Thus, an automatic classifier likely only needs to be trained on a select few units because they will be repeated often enough for the classifier to inform a decision of what cetacean species is present.

## 3.2 BOWHEAD WHALE

The bowhead whale (*Balaena mysticetus*) is a baleen whale with a nearly circumpolar distribution in Arctic and sub-Arctic waters. The Eastern Arctic population of this large whale is listed on the Canadian SARA list as "Endangered," although other Arctic populations have either no status or are listed as "Special Concern;" populations of these whales were severely depleted due to excessive whaling. Since bowheads spend significant amounts of time close to the edge of the pack ice, climatic conditions that influence ice conditions may significantly affect the survival and distribution of the species. Additionally, elevated levels of ship traffic in the Arctic lead to increased chance of mortality due to ship strike [30].

The bowhead vocalizations used for this research came from the MobySound website [32]. A significant advantage of using the bowhead MobySound dataset is that experts have previously classified the data with text annotations indicating detections. The data containing bowhead vocalizations was recorded off of Point Barrow, Alaska using homemade hydrophones with Sippican transducer elements – sampling was performed at a rate of 4.0 kHz. The MobySound archive contained only the endnotes of the bowhead song and none of the other parts of the species' complex song – bowhead songs change from year to year, although the endnotes remain relatively constant and tend to exhibit higher source levels than other parts of the song [31], [33] and are thus the most accepted song component for classification purposes.

Classification was performed on the bowhead whales' song endnote (Figure 3.3). The spectrogram of these sounds forms a distinctive undulating shape, with frequencies ranging between about 50 – 800 Hz. The duration of these sounds is approximately 2.5 – 3.0 s [31]. There were 259 bowhead vocalizations used for classification.

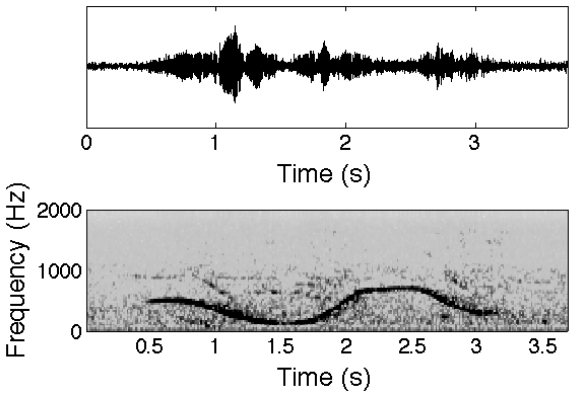


Figure 3.3 Time series and spectrogram of a bowhead song endnote. The spectrogram was generated using a Hamming window length of 256 samples and 60% overlap.

Bowhead and humpback vocalizations have very similar frequency bandwidth and duration, so that many automatic detection/classification schemes have difficulty distinguishing between these two species. A primary reason for selecting bowhead and humpback whales for this classification work was to provide a case that has challenged many automatic classification algorithms in the past.

## 3.3 HUMPBACK WHALE

Depending on the time of year, humpback whales (*Megaptera novaeangliae*) can be found worldwide in tropical, temperate, and sub-polar waters. These baleen whales migrate seasonally between high-latitude summer feeding grounds, and low-latitude breeding and calving areas in the winter. The Atlantic humpback population is listed as "Special concern" and the Pacific population is listed as "Threatened" on the Canadian SARA list [30]. Humpbacks are the second-most commonly reported marine animals to be the subject of ship strikes (following fin whales) [36]; since they are such a vocal

species, passive acoustic detection and classification may be used to reduce the number of ship strikes involving humpback whales.

Male humpback whales sing primarily in the winter while on the breeding grounds, although some singing whales have been noted during migration and while on summer feeding grounds. Humpback song changes significantly from year to year with complete change noted within two to five years [31]. Humpbacks in different ocean basins (e.g. Atlantic and Pacific populations) have different song dialects; that is, distinct populations sing different songs that do not appear to have any commonalities other than their hierarchical structure, whereas all whales within the same population sing the current version of that population's song [37]. Because the unit is the basic song component, it is believed that there will be less variation in units from year to year within a population and that there may be some common units between populations. Therefore, only distinct song units were selected for classification and no attempt was made to include longer components like phrases. Due to the broad vocal repertoire of humpback whales, there is overlap in frequency range and duration with many other marine mammals including North Atlantic right whales and bowhead whales. Because of this overlap with vocalizations from other species, humpback whale sounds may be a confounding factor when monitoring other species [38].

The humpback vocalizations were obtained from the MobySound website. Recordings of humpback song were made off the north coast of the island of Kauai, Hawaii using custom-built hydrophones with a Sippican transducer element [32]. Data was provided with a sampling frequency of 4.0 kHz.

Of the many different sounds produced during humpback whale song, four song units (Figure 3.4) were selected for the purposes of this project. These particular units were selected to provide overlap in frequency content and duration with both bowhead and North Atlantic right whale vocalizations to provide a challenging case to test the robustness of the aural classifier. These units occurred relatively frequently throughout the recordings, so in a realistic setting the classifier would not need to classify every

sound that a humpback made but instead it would be sufficient to recognize one of these units as belonging to a humpback whale.

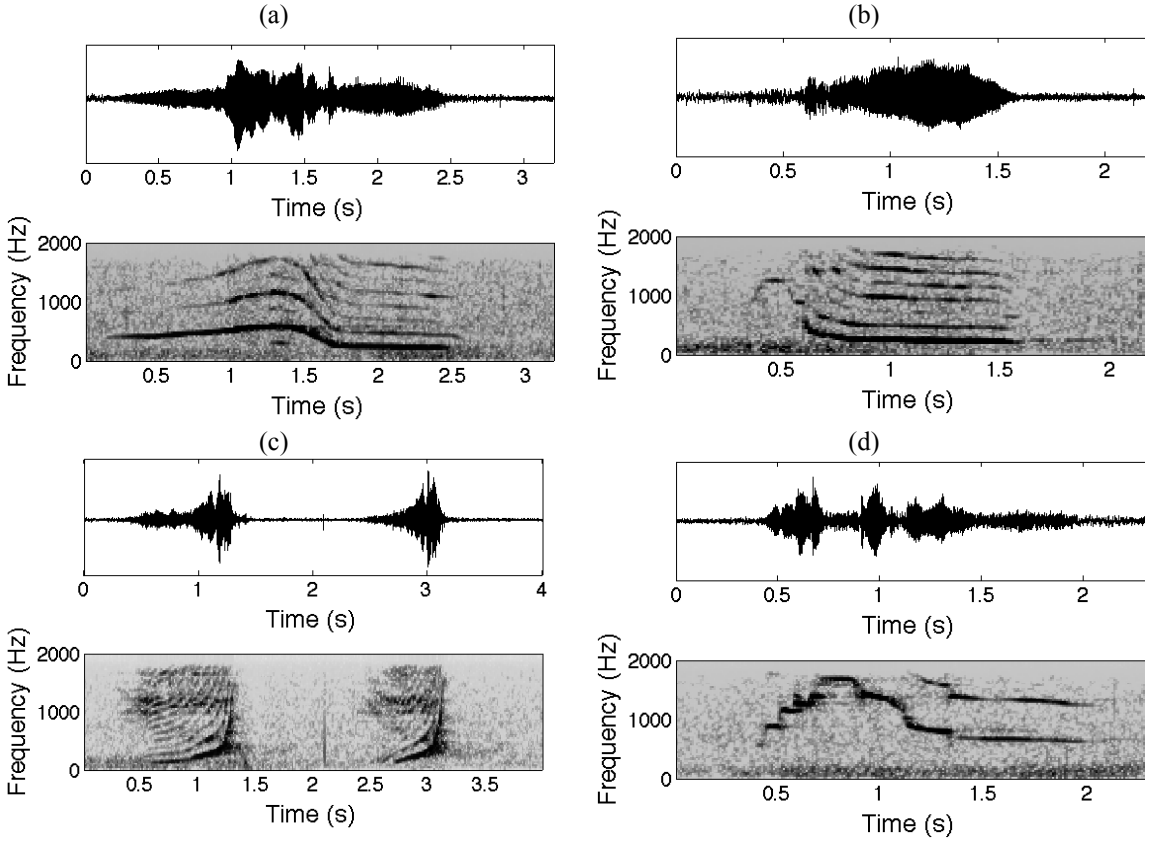


Figure 3.4 Time series and spectrograms of the four humpback units selected for classification. Spectrograms were generated using a Hamming window length of 256 samples with 60% overlap. These units will be referred to as (a) humpback1, (b) humpback2, (c) humpback3, and (d) humpback4. In (c) two humpback3 units are shown – during classification only a single unit of this type would be used (i.e. unit is less than one second in length).

The overtone structure of each of these units is very apparent. Table 3.2 lists the frequency bandwidth and duration of the fundamental frequencies for each of the selected units. Each instance of the humpback1 and humpback4 units exhibited enough aural similarity to be recognized as belonging to the respective subclass although the frequency extent and spectrogram contour showed substantial variation between instances – spectrogram correlation techniques used by many researchers would not work well with these units although because of the aural similarities of these units it is expected that aural classifier results will not be significantly affected. There were a total of 456

humpback units available for classification; of these there were 206 humpback1 units, 122 humpback2 units, 83 humpback3 units, and 45 humpback4 units.

Table 3.2 Approximate frequency bandwidth of the fundamental frequency and duration of the four types of humpback units shown in Figure 3.4.

| | **Frequency Bandwidth (Hz)** | **Duration (s)** |
|---|---|---|
| Humpback1 | 200 – 600 | 2.5 – 3.0 |
| Humpback2 | 150 – 700 | 1.0 – 1.5 |
| Humpback3 | 100 – 2000 | ~ 1.0 |
| Humpback4 | 500 – 1300 | 1.5 – 2.0 |

## 3.4 NORTH ATLANTIC RIGHT WHALE

North Atlantic right whales (*Eubalaena glacialis*), a baleen whale species, are currently listed as "Endangered" on the Canadian SARA list with estimates of 300 – 400 individuals remaining and are generally considered to be among the most endangered whales in the world. The right whales are particularly at risk of ship strikes or becoming entangled in fishing gear, in part because they spend large amounts of time close to the surface and because they migrate close to shore in areas of high ship traffic. The eastern stock of North Atlantic right whales have been sighted in coastal waters between the Canary Islands and Norway, whereas the western North Atlantic stock spends winters off the coasts of Florida and Georgia and summers off the north-eastern United States and Canada (see figure Figure 3.5) [30]. Anthropogenic mortalities are currently responsible for about 40% of known right whale mortalities. Population projections performed in 2001 suggest that, given current mortality rates, the North Atlantic right whale population will be extinct within 100 – 400 years [15]. The future of the right whale depends on a significant reduction in deaths caused by human activities [39].

North Atlantic right whale vocalizations were recorded in the Bay of Fundy by previous effort of DRDC scientists. Data was collected using a variety of sonobuoy types and a CP140 Maritime Patrol Aircraft. Data was stored with a sampling frequency of 6.554 kHz.

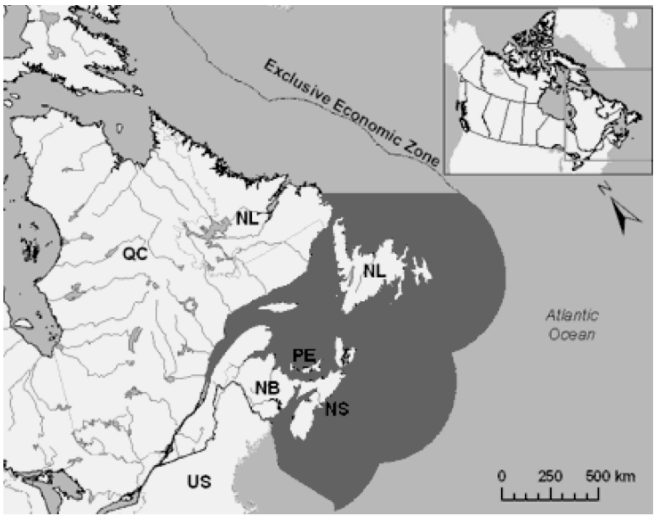


Figure 3.5 Known distribution of North Atlantic right whales in Canadian waters, shown in dark grey [30].

Two general types of sounds are considered for right whale classification – frequency modulated sounds that are at least one second in length, referred to here as cries and moans, and the distinctive "gunshot" sound that is a short broadband transient with duration less than 0.5 s. Examples of the North Atlantic right whale sounds used for classification are shown in Figure 3.6. There is significant variability in how the moan-like calls' frequency contour changes with time, as well as differences in durations (ranging from 1.0 – 2.5 s); however, these calls have a common range of fundamental frequencies (70 – 190 Hz). The cries are 1.0 – 1.5 s in duration and occur in the 400 – 500 Hz frequency band. Some overtones are evident in the right whale moan and cry sounds. The example gunshot sound has at least three noticeable pulses – the initial impulse is the original right whale gunshot sound and the following impulses are likely due to multipath reflections.

In total there were 142 North Atlantic right whale vocalizations used for this research; these can be further divided into the subclasses shown in Figure 3.6. There were 26 moan-like sounds, 30 cry sounds, and 86 gunshot sounds.

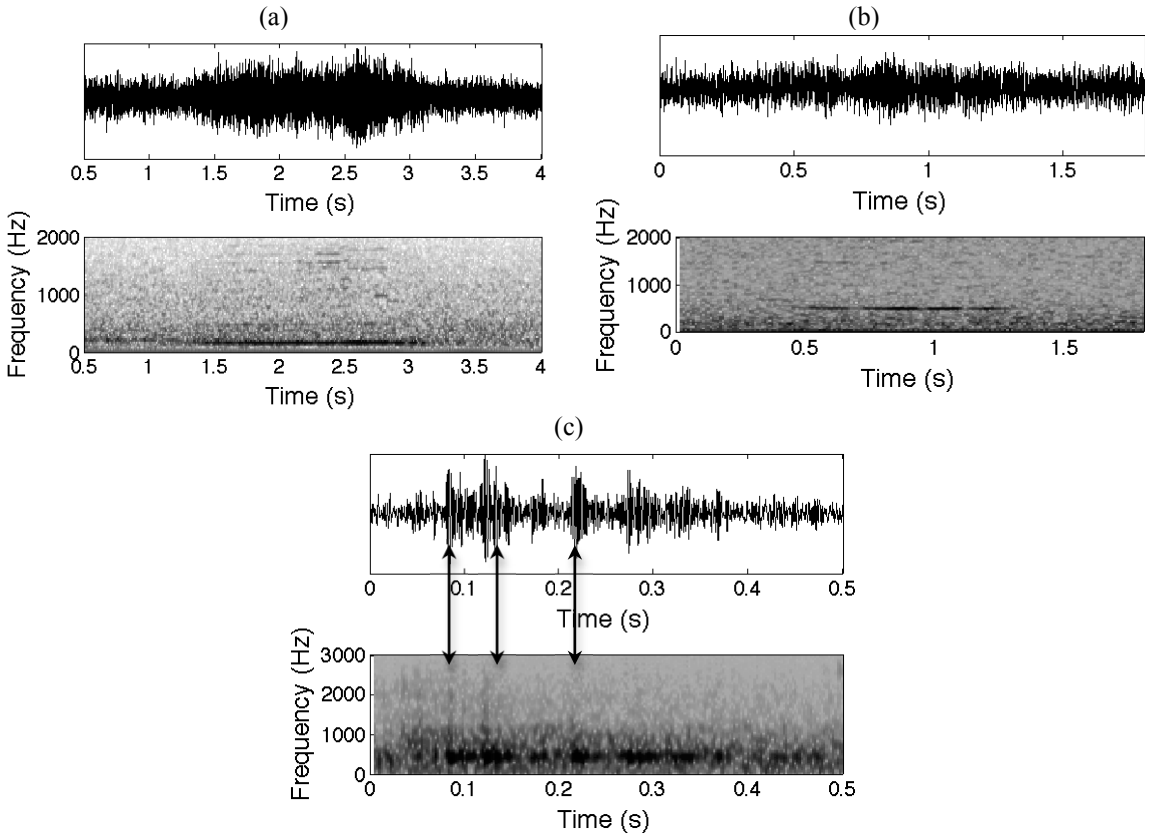


Figure 3.6 Time series and spectrograms for (a) right whale moan, (b) cry, and (c) gunshot sounds. Spectrograms were generated using a Hamming window length of 256 samples and overlap of 70% for the moan and cry sounds. The right whale gunshot spectrogram was generated using a Hamming window size of 64 samples and 70% overlap. The arrows on (c) indicate the locations of the three impulses associated with the gunshot sound.

## 3.5 SPERM WHALE

Sperm whales (*Physeter macrocephalus*) are one of the most difficult large whale species to detect using traditional visual observation techniques because they live off of continental shelves worldwide, perform deep dives and spend approximately only ten minutes at the surface [40] between dives, which can last up to 90 minutes [10]. Sperm whales are the only odontocete species included in the dataset. This whale species generates powerful, distinct sounding clicks that can be detected at ranges of several kilometres, which can be used to acoustically detect this species [40]. The primary reason for including sperm whale clicks in this dataset was because of the similarity of

the click structure with a wide variety of anthropogenic noise sources (i.e. false alarms) that could confuse many different types of automatic classifiers.

Data collected during a Canadian Forces Auxiliary Vessel (CFAV) QUEST– DRDC's research ship – research cruise contained recordings of many sperm whale clicks. These clicks were collected by a SSQ57B broadband sonobouy with sampling frequency of 80 kHz. Recordings were collected at the Atlantic Undersea Test and Evaluation Center Range which is a location that sperm whales are common – thus, contextual information helped confirm the initial classification of acoustic data [31].

Sperm whale clicks are very short in duration (as can be seen in Figure 3.7), about 2 – 3 ms in length and often contain multiple arrivals spaced several hundredths of seconds apart. It was determined that the double click phenomenon shown is likely due to multipath effects rather than a multi-pulse structure caused at the source, since the spacing between the two clicks is consistent with that attributed by Thode *et al.* to result from a direct path and bottom reflection [41]. Sperm whale clicks cover a frequency range of about 500 Hz – 17 kHz [31], but in this case detection and classification is limited to 0 – 4 kHz because of data re-sampling. There were 178 sperm whale clicks detected and used for classification.

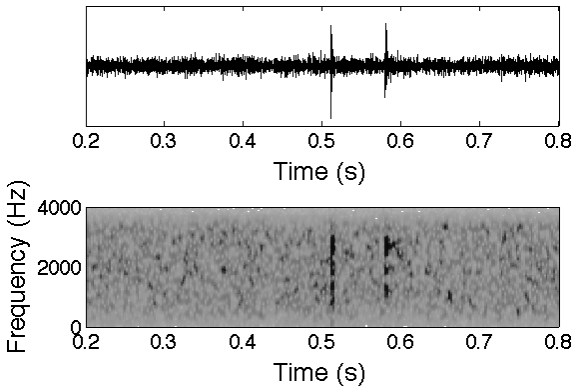


Figure 3.7 Time series and spectrogram of a sperm whale click. The spectrogram was generated using a Hamming window length of 64 samples and an overlap of 70%.

## 3.6 MINKE WHALE

Minke whales (*Balenoptera acutorostrata*) are very difficult to detect visually, especially in rough sea conditions. Minke whales are relatively common and have a worldwide distribution, but sightings of this species are rare because they are the smallest of the baleen whales, are encountered individually or in small groups of two or three, have relatively inconspicuous blows and only spend small amounts of time at the surface [42]. Because of the low probability of visual sightings, minke whales are ideal candidates for studying using passive acoustic methods.

North Pacific minke whales seasonally (November – March) generate a unique sound known as the "boing". This sound has recently been attributed to the minke whale using visual and passive acoustic methods [42] and has since become an accepted cue for detection and classification purposes. Boings produced by the Hawaiian minke whale population have a mean duration of 2.6 s [29]. The duration and frequency content of boings are relatively constant from year to year, although there are some variations based on population (i.e. spatial variation) [42]. False alarms caused by humpback song have occurred during automatic detection of minke boings because of some overlap in the frequency content and duration of some humpback units and the minke whale boing [29].

Figure 3.8 shows an example of a minke whale boing; note the initial brief pulse and subsequent frequency- and amplitude-modulated long call. Boings usually have several overtones apparent in their spectrograms. This project made use of a dataset containing minke boing recordings that was released for testing various automatic algorithms as part of the 5$^{th}$ International Workshop on the Detection, Classification and Localization Using Passive Acoustics [32], [43].

The boings were recorded at a US Navy test range off the coast of Kauai, Hawaii using seven bottom-mounted hydrophones. The data-sampling rate was 96 kHz with 16-bit resolution. From this dataset, a total of 127 minke boings were extracted for classification.

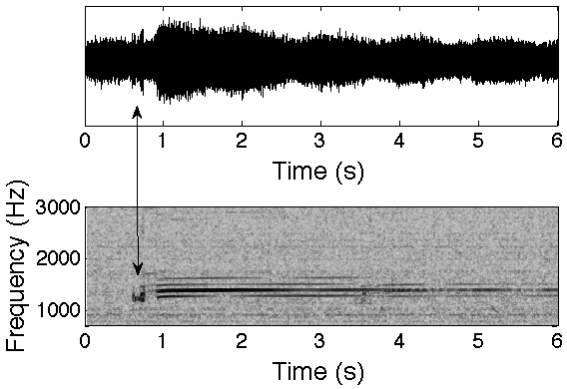


Figure 3.8 Time series and spectrogram of the minke "boing" sound. The spectrogram was generated using a Hamming window length of 2048 samples with an overlap of 75%. The arrow indicates the location of the initial brief impulse.

## 3.7 DATASET SUMMARY

The following table summarizes the total number of vocalizations, per species, contained in the cetacean dataset. The "Species Total" column represents the sum of the separate vocalization units for a species, e.g. there are a total of 142 North Atlantic right whale vocalizations when all types of sounds are considered.

Table 3.3 Number of vocalizations, by species, in the cetacean dataset. The number of vocalizations is broken down by units where applicable.

| **Species** | **Number of Vocalizations** | **Species Total** |
|---|---|---|
| Bowhead | 259 | |
| Humpback1 | 206 | 456 |
| Humpback2 | 122 | |
| Humpback3 | 83 | |
| Humpback4 | 45 | |
| North Atlantic right moans and cries | 56 | 142 |
| North Atlantic right gunshots | 86 | |
| Minke | 127 | |
| Sperm | 178 | |
| **Total Number of Vocalizations** | 1162 | |

# CHAPTER 4 DATA PREPARATION AND DETECTION PROCESS

## 4.1 DATA PREPARATION

Acoustic data pre-processing and automatic detection of vocalizations were performed by Akoostix Inc. of Dartmouth, NS as outlined in Ref. [31]. Various sampling strategies were used during initial recordings – to ensure consistent detection processing and classification, all data were re-sampled to 8.0 kHz, using the open source Linux sound exchange (SoX) application [44], with quadratic interpolation.

This required a severe down-sampling of the sperm whale data – it was aurally confirmed that enough information was left available after down-sampling for automatic aural classification purposes. The reduced sampling rate is also consistent with using reduced bandwidth on the original recording system, and is a realistic scenario in that context.

## 4.2 DETECTION PROCESS

The acoustic recordings obtained from the various sources contained many individual vocalizations in data files ranging from a couple of minutes to approximately half an hour long. Aural classification is performed on a single vocalization – so individual vocalizations must first be located within the recordings. Using a process called band-limited energy detection, Akoostix Inc. provided vocalizations to be used for this research. A variety of vocalizations were desired – including potentially challenging vocalizations (i.e. relatively low SNR) – to evaluate the robustness of the aural classifier.

Band-limited energy detection is a common technique for detection of marine mammal vocalizations, because many of the signals have a characteristic bandwidth and duration, but too much variability for correlation-based techniques [45]. For example, the humpback whale is known to change its song from year to year so correlation methods will not work over a longer time period, but there may be sufficient similarities in the duration and frequency content of the units that a band-limited energy detector will still function well. In performing energy detection, a detection function is calculated by estimating the short-term average energy in the signal band and dividing it by a longer average of the background noise energy; this forms the basis for a likelihood ratio test. The value of the computed ratio is then compared to a pre-defined acceptable threshold value [45].

The automatic detector used for this research offers two different methods to estimate energy content in the signal band. The first method, typically used for impulsive sounds (like sperm whale clicks), uses an exponential average of the form,

$$y[n] = \alpha y[n-1] + (1-\alpha)x[n] \quad , \qquad \text{Eqn. 4.1}$$

where $x$[n] is the energy of the $n^{\text{th}}$ sample, and $y[n]$ is the energy estimate for sample $n$. The averaging coefficient, $\alpha$, is defined by,

$$T_c = \frac{\Delta T}{1-\alpha} \quad , \qquad \text{Eqn. 4.2}$$

where $T_c$ is a time constant and $\Delta T$ is the time resolution. The value of $\alpha$ ranges between zero and one [31], [45]. The second method, more efficient for detection of narrower band signals (like baleen whale vocalizations) employs a split window to estimate the energy average. The split window method uses two rectangular windows to provide estimates for the signal and noise levels. Windows are defined such that the noise window ($W_N$) is longer than the signal window ($W_S$) and each window has an odd number of samples so that there is no ambiguity about the location of the window centre. The estimate for the signal level at sample $n$ is given by,

$$y_S[n] = \frac{1}{W_S} \sum_{i=-\frac{1}{2}(W_S-1)}^{\frac{1}{2}(W_S-1)} x[n+i] \quad . \qquad \text{Eqn. 4.3}$$

The noise estimate is found in a similar way using,

$$y_N = \frac{1}{W_N - W_S} \sum_{i=-\frac{1}{2}(W_N-1)}^{\frac{1}{2}(W_N-1)} \left( x[n+i] - y_s[n] W_S \right) \quad , \; W_N > W_S . \qquad \text{Eqn. 4.4}$$

The level estimate provided by the split window method is non-causal, so that in practice the output of the detector will lag behind the signal input by approximately half the noise window length. The estimates of signal and noise energy are used to perform the likelihood ratio test – if the likelihood ratio exceeds the pre-defined threshold value, a detection is generated. For this research, the threshold value was set relatively low to include vocalizations with a range of SNR values.

Parameters used for the automatic detection process are listed in Table 4.1. Frequency bands were selected to include most of the energy from the first harmonic of the vocalization. The split window estimation method was used for all types of whale vocalizations, except for the sperm whale clicks. Sperm whale clicks were detected using the exponential average method. Humpback3 units were not specifically configured as detection targets, but instead resulted from detection parameters for the other humpback units; it was decided to include the humpback3 units because of the large number of detections and similarity in frequency bandwidth and duration to vocalizations of bowhead and right whales. Detection parameters were selected to allow as many detections as possible, while also generating relatively large numbers of false alarms. The philosophy is that the classification process will significantly reduce the false alarm rate and correctly identify many of the detections; if more stringent detection parameters were used, any missed detections would remain unclassified [31].

Table 4.1 Detection parameters used for each type of cetacean vocalization. The listed parameters define the signal band.

| Whale Vocalization Type | Frequency Resolution (Hz) | Time Resolution (s) | Low Frequency (Hz) | High Frequency (Hz) | Signal Window Size (s) |
|---|---|---|---|---|---|
| Bowhead | 5 | 0.1 | 50 | 700 | 1.0 |
| Humpback1&2 | 5 | 0.1 | 200 | 500 | 2.0 |
| Humpback4 | 5 | 0.1 | 625 | 1550 | 2.0 |
| Right Moan | 5 | 0.1 | 120 | 220 | 1.5 |
| Right Cry | 5 | 0.1 | 415 | 515 | 1.5 |
| Right Shot | 5 | 0.1 | 1325 | 1425 | 1.5 |

| Whale Vocalization Type | Frequency Resolution (Hz) | Time Resolution (s) | Low Frequency (Hz) | High Frequency (Hz) | Signal Window Size (s) |
|---|---|---|---|---|---|
| Minke | 0.73 | 0.505 | 1100 | 1600 | 1.0 |
| Sperm | 100 | 0.005 | 1000 | 3900 | **Time Constant** = 0.005 |

# CHAPTER 5 CETACEAN CLASSIFICATION

This chapter presents aural classification results using the cetacean dataset described in CHAPTER 3. Aural classification is accomplished using the process outlined in Sections 2.1.1 to 2.1.5 – principal component analysis is used to project the data onto the two-dimensional space. The same training and testing subsets were used for all results presented in this chapter.

## 5.1 MULTICLASS CLASSIFICATION

### 5.1.1 All Cetacean Species ($c$ = 5)

The first analysis examines a multiclass classification that includes all five species in the dataset. Fourteen features were removed during redundancy reduction, which left 44 features for consideration. As discussed in Section 2.1.4, it is unlikely that all the perceptual features will be useful in discriminating between the cetacean species' vocalizations; thus, a subset of features is selected prior to performing PCA. To determine the number of features to be used in PCA, a plot of the cumulative Fisher score with respect to number of features (Figure 5.1) is used. The cumulative Fisher score is determined by summing the Fisher scores ($s_D$) of the first $k$ features, $\sum_{i=1}^{k} s_{d,i}$, where the features have been sorted in order of decreasing Fisher score. This type of plot represents the relationship between number of selected features and discriminability. A bend in the plot, where the curvature of the plot begins to noticeably decrease, indicates the point at which including additional features will not significantly increase classifier performance. The features corresponding to the flat portion of the curve have low Fisher scores because

they do not do as well at discriminating between classes – including these features will not significantly add to classifier performance and may, in fact, decrease classifier performance by introducing noisy features.

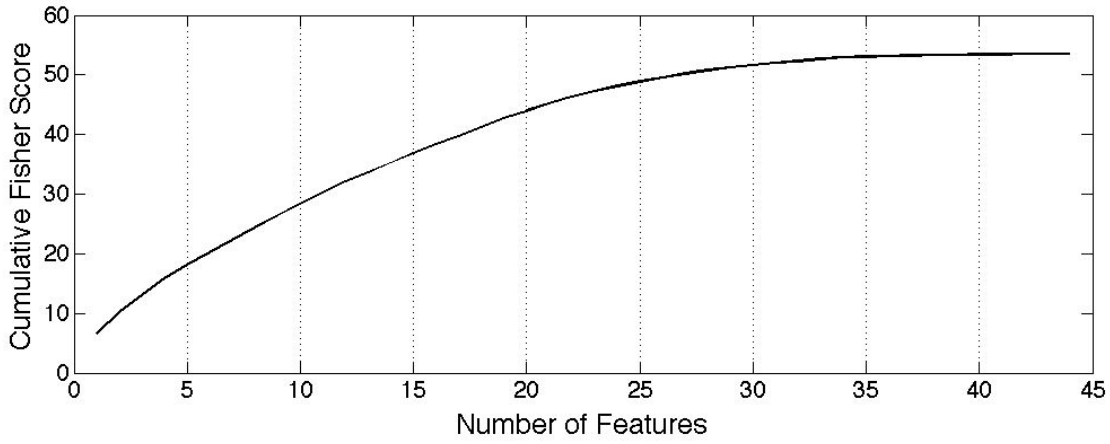


Figure 5.1 Cumulative Fisher score with respect to number of features for all cetacean species. Note that the features have been sorted so that the first feature has the largest Fisher score. Points are connected for visualization purposes and not intended to imply the data are continuous.

Figure 5.1 represents the cumulative Fisher score for the perceptual features calculated when all five cetacean species are considered. The Fisher score curve is smoothly varying with no obvious bend; there is a relatively constant increase in the cumulative Fisher score between 1 and 30 features. At 30 features the cumulative Fisher score curve is relatively flat. Closely inspecting the curve reveals a slight decrease in the slope of the curve at 20 features that becomes more noticeable at larger number of features. Based on these results, classification will be performed using 30 and 20 features.

Classification of all five cetacean species was first performed using 30 features, with the particular classifier model derived from the training subset. The decision region obtained from the *training* subset is shown in Figure 5.2a with classification accuracy of 90%. The confusion matrix of pairwise *AUC* values is shown in Table 5.1; the average of these pairwise *AUC* values gives an *M*-measure of 0.99, which is indicative of a successful classification over all classes. The decision region with the data from the *testing* subset is shown in Figure 5.2b. Observe how the decision boundaries are the same for both

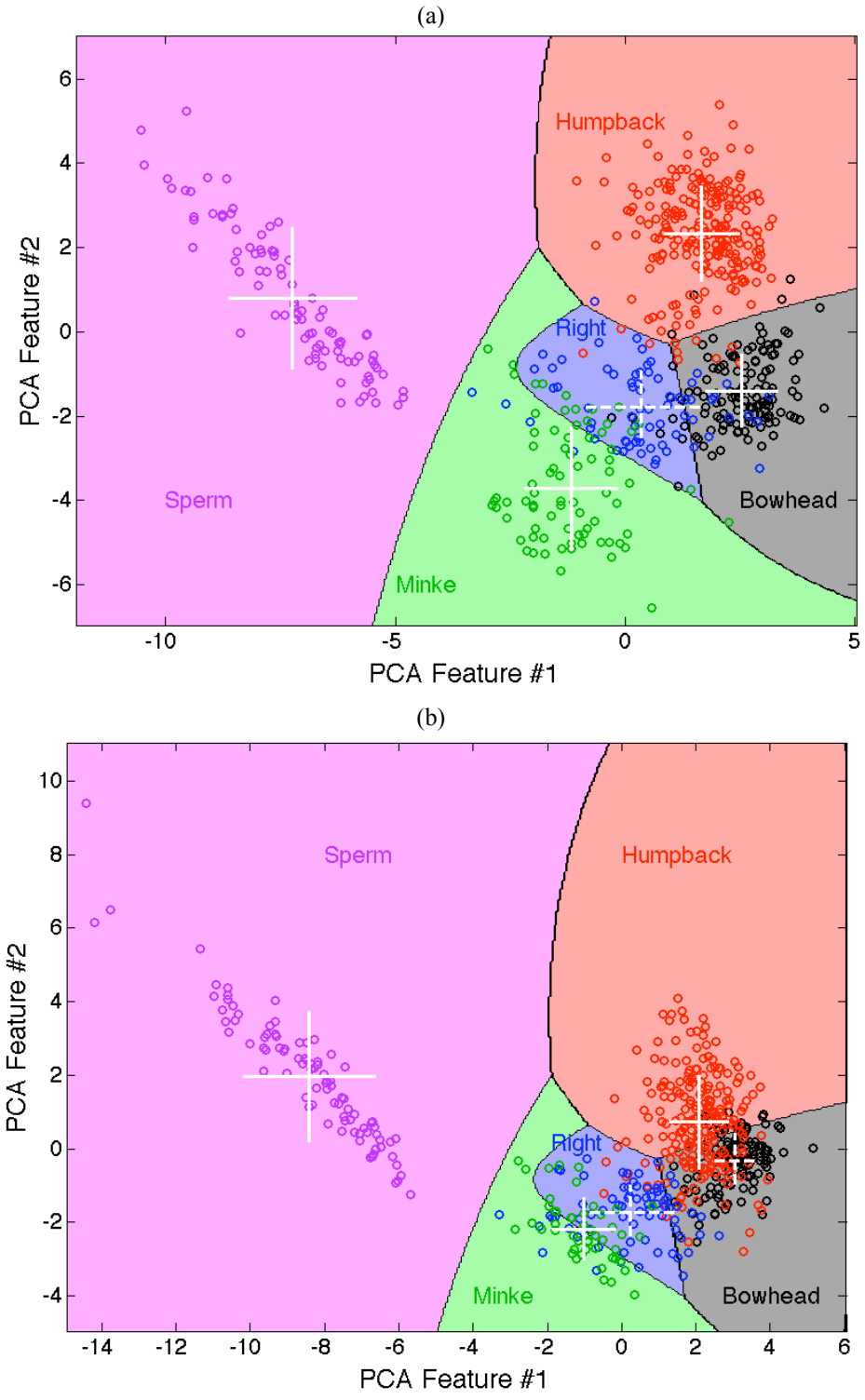


Figure 5.2 Decision regions for multiclass classification of all cetacean vocalizations. (a) Results from the *training* subset and (b) results from the *testing* subset. Classification was performed with 30 selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

decision regions – this is because only the data from the training subset are used to calculate the decision boundaries, the results of which are then applied to the testing subset. Remember that the decision regions are obtained by fitting Gaussian PDFs to each class (using data in the training subset) and the boundaries between regions are defined by points of equal likelihood probability. Results from the training subset are shown here to provide an example of how the decision region relates to the data in the training subset. For all other circumstances, only the results from the testing subset will be presented because they convey information about how successfully the classifier model can be applied to a different dataset.

Table 5.1 Confusion matrix of *AUC* values corresponding to the decision region shown in Figure 5.2a. The value $M = 0.99$. The asterisk indicates *AUC* values that appear to result from an ideal classifier, but in fact resulted from rounding to two decimal places.

| | **Humpback** | **Right** | **Minke** | **Sperm** |
|---|---|---|---|---|
| **Bowhead** | 1.00* | 0.94 | 1.00* | 1.00 |
| **Humpback** | | 1.00* | 1.00 | 1.00 |
| **Right** | | | 0.95 | 1.00 |
| **Minke** | | | | 1.00 |

The decision region for classification of all five cetacean species in the testing subset, using 30 selected features, is shown in Figure 5.2b with a corresponding *M* value of 0.97 – indicative of a successful classification over all classes. The confusion matrix of pairwise *AUC* values is shown in Table 5.2. The vocalizations in the test set were classified with 75% accuracy. All sperm whale clicks were correctly classified as can be seen both by examining the decision region and from the fact that the column representing the pairwise classifications with sperm whales contains *AUC* values of 1.00 for all cases. The most overlap occurred between the humpback/bowhead and minke/right whale pairs of classes. When the first two principal components were used a value of $p_2 = 0.59$ was obtained, indicating that more than half of the variance contained in the 30-dimensional feature space is represented in this reduced 2D space. The sperm whale class contains the largest amount of within-class variance, while the within-class variances of the other four classes are similar to one another, as can be noted by the size

Table 5.2 Confusion matrix of *AUC* values corresponding to the decision region shown in Figure 5.2b. The value $M = 0.97$. The asterisk indicates *AUC* values that appear to result from an ideal classifier, but in fact resulted from rounding to two decimal places.

| | **Humpback** | **Right** | **Minke** | **Sperm** |
|---|---|---|---|---|
| **Bowhead** | 0.83 | 0.99 | 1.00 | 1.00 |
| **Humpback** | | 0.98 | 1.00* | 1.00 |
| **Right** | | | 0.86 | 1.00 |
| **Minke** | | | | 1.00 |

of the white crosses in Figure 5.2b. An interesting feature of this plot is the linear spread of sperm whale points, especially compared with the more random spread of data points in the other classes – this trend will be discussed in Section 5.3.2.

The multiclass classification was repeated with only 20 features selected. The corresponding decision region is depicted in Figure 5.3 and the matrix of pairwise *AUC* values is represented in Table 5.3. The results of the classifier were good with

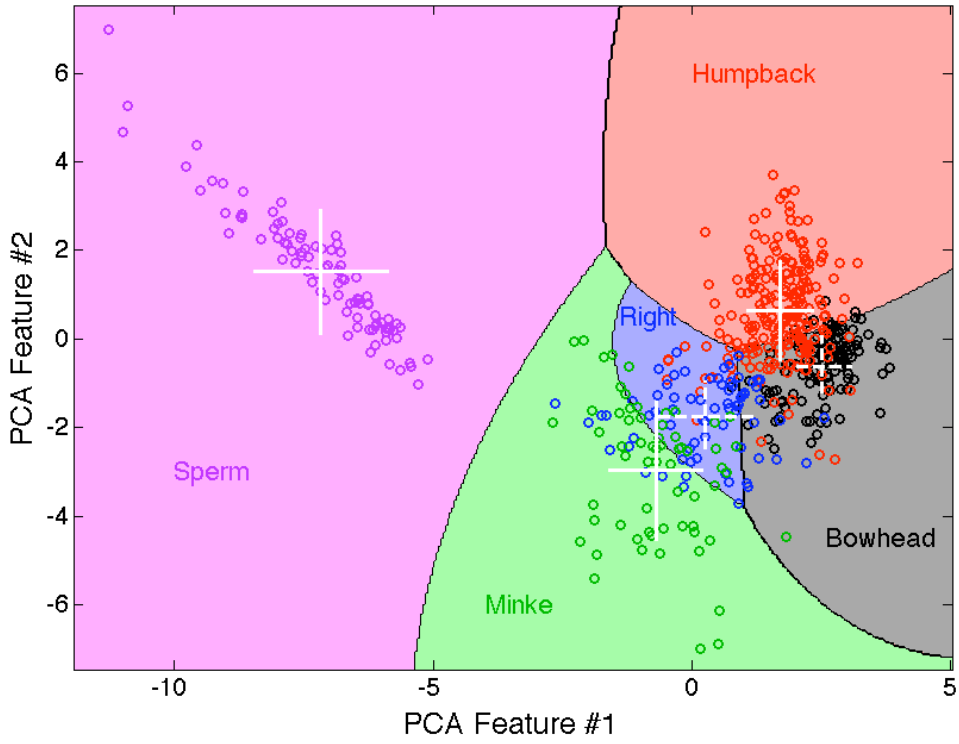


Figure 5.3 Decision region for multiclass classification of all cetacean vocalizations. Classification was performed with 20 selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

$M = 0.97$ and 78% accuracy. These classification results are similar to classification humpback/bowhead and minke/right whale classes as indicated by the larger $AUC$ values for these pairs of classes. The minke whale class has a large amount of within-class variance, as represented by the size of the white cross in Figure 5.3; based on *ad hoc* listening tests performed by the author, this was not expected because there seems to be little variation in how the minke whale vocalizations sound. Once again, the sperm whale data points were distributed linearly. More of the variance in the selected features is maintained in this case where $p_2 = 0.67$. The fact that there were ten fewer features to contribute to the variance of the higher dimensional feature space likely contributed to being able to maintain more variance in the principal components when only 20 features were selected.

Table 5.3 Confusion matrix of $AUC$ values corresponding to the decision region shown in Figure 5.3. The value $M = 0.97$. The asterisk indicates $AUC$ values that appear to result from an ideal classifier, but in fact resulted from rounding to two decimal places.

| | **Humpback** | **Right** | **Minke** | **Sperm** |
|---|---|---|---|---|
| **Bowhead** | 0.89 | 0.99 | 1.00 | 1.00 |
| **Humpback** | | 0.98 | 1.00* | 1.00 |
| **Right** | | | 0.87 | 1.00 |
| **Minke** | | | | 1.00 |

A prominent feature in both decision regions (Figure 5.2b and Figure 5.3) is the apparent linear trend of the sperm whale click data points. When Gaussian PDFs are fit to each class, it is assumed that the within-class covariance is negligible (i.e. approximately zero). In the case of the sperm whale clicks this is not true – the covariance and correlation, in the case of thirty selected features are -3.02 and -0.95, respectively. When twenty features were selected the covariance was -1.75 with a corresponding correlation of -0.95. This would indicate that the two principal components are highly correlated for sperm whale clicks and the assumption that the off-diagonal components of the covariance matrix are zero is not valid. A possible cause for the non-zero within-class covariance is presented in Section 5.3.2.

The three most important features in the PCA space are listed in Table 5.4. Examining the eigenvectors that define the PCA transformation (as in Figure 5.4) exposes the relative weighting of each feature used during projection onto the reduced feature space. The absolute value of the eigenvectors are summed and normalized by the maximum feature weight and plotted to provide a simple method for identifying the relative importance of each feature. For example, the normalized sum of eigenvector components corresponding to the peak loudness frequency feature (feature number 21 on the plot) for the principal components composed of 30 features is 0.62 compared to 0.58 for the 20-feature principal components; thus, it can be concluded that peak loudness frequency is slightly more important when 30 features are selected for the PCA method.

Table 5.4 Three highest weighted features using 30 and 20 selected features for classification of the five cetacean species.

| **30 Selected Features** | **20 Selected Features** |
|---|---|
| Integrated loudness | Integrated loudness |
| Psychoacoustic bin-to-bin difference | Psychoacoustic bin-to-bin difference |
| Pre-attack psychoacoustic maxima-to-spectral-bins ratio | Pre-attack peak loudness value |

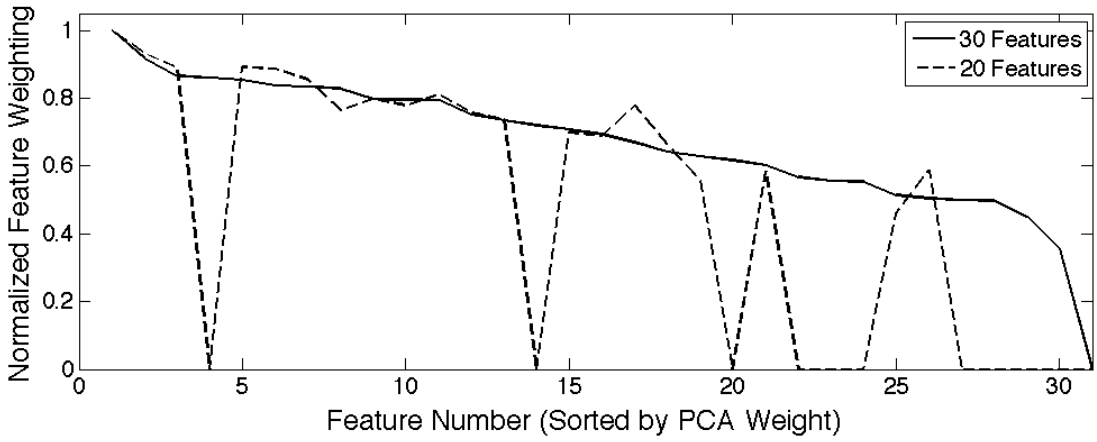


Figure 5.4 Normalized weighting of features in the first two principal components. Features are sorted from largest PCA feature weighting to smallest based on PCA with 30 selected features. These eigenvectors correspond to the decision regions shown in Figure 5.2b and Figure 5.3. Peaks are connected merely for visualization purposes and are not intended to imply that the data are continuous.

Relative feature weightings within the first two principal components for multiclass classification of all five cetacean species are presented in Figure 5.4. Two of the top three features are the same when either 20 or 30 features were selected. The same general trend in feature weighting is observed for both 20 and 30 selected features. The features that received zero weighting correspond to features that were not included in PCA when only 20 features were selected. The Fisher Linear Discriminant score is used to select features that will best separate classes from each other, whereas PCA best maintains the variance in the entire dataset. Because the method for selecting features has a different goal than PCA, the features with zero weighting do not necessarily correspond to the features with the lowest weighting when PCA is performed with 30 features.

### 5.1.2 Baleen Species ($c$ = 4)

When classification was performed with all species, the sperm whale clicks separated out very well with no misclassifications; however, there was overlap between the baleen whale classes. The sperm whale clicks separated out so well because they sound so distinct, whereas the baleen whale vocalizations share similar aural characteristics. This section examines classification of only the baleen whale species – bowhead, humpback, right and minke whales – to determine which aural features best discriminate between these four species.

The plot of cumulative Fisher score, shown in Figure 5.5, exhibits a bend at 8 features and the slope of the curve becomes small around 25 features. Based on these trends in the cumulative Fisher score, classification was performed with both 8 and 25 features to determine if classification results improve by including more features.

The decision region whose axes are combinations of 25 selected features is shown in Figure 5.6. This classification resulted in a total accuracy of 69%. The pairwise *AUC* values corresponding to this decision region are listed in Table 5.5 and average together to give $M = 0.94$. The aural classifier did a good job of discriminating minke whale

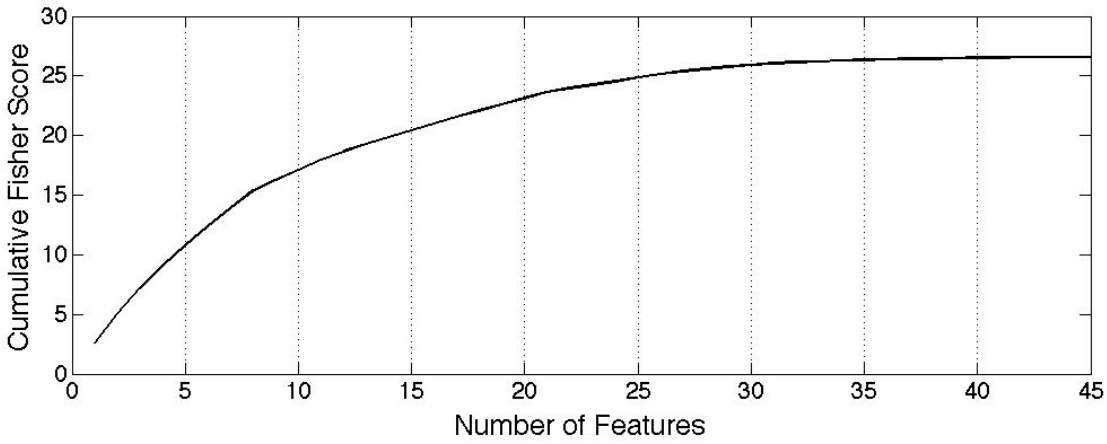


Figure 5.5 Cumulative Fisher score with respect to number of features for baleen whale species. Note that the features have been sorted so that the first feature has the largest Fisher score. Points are connected for visualization purposes and not intended to imply the data are continuous.

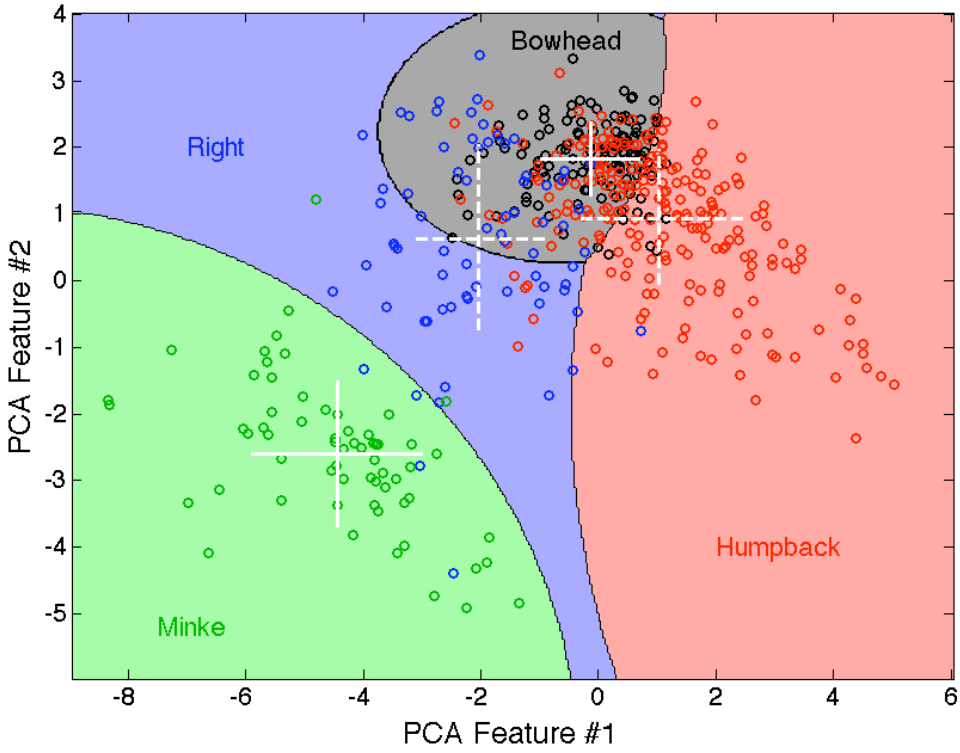


Figure 5.6 Decision region for multiclass classification of baleen whale vocalizations. Classification was performed with 25 selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

Table 5.5 Confusion matrix of *AUC* values corresponding the to decision region shown in Figure 5.6. The value $M = 0.94$.

| | **Humpback** | **Right** | **Minke** |
|---|---|---|---|
| **Bowhead** | 0.80 | 0.88 | 1.00 |
| **Humpback** | | 0.95 | 1.00 |
| **Right** | | | 0.99 |

vocalizations – no humpback/minke or bowhead/minke misclassifications were made. There was a small amount of overlap between the minke and right whale classes. The most overlap occurred between the humpback/bowhead and bowhead/right whale classes. The bowhead class had the least amount of within-class variance. Less than half the total variance in the 25 selected features was captured in the first two principal components ($p_2 = 0.49$).

Classification of the four baleen whale species was also performed with only eight selected features. The decision region (Figure 5.7) and confusion matrix of pairwise *AUC* values (Table 5.6) present the classification results. The multiclass performance measure, $M = 0.96$, and accuracy of 79% were indicative of successful classification over all classes. Both the decision region and pairwise *AUC* values present results consistent with high variance in the minke whale class, as indicated by the white cross on the decision region – the length of the arms of the white cross representing the standard deviation of the minke whale dataset are longer than for any of the other classes – this resulted in relatively large overlap with the other three classes. When eight features were selected $p_2 = 0.71$, indicating that a large amount of the variance present in the selected features is maintained in the first two principal components.

Table 5.6 Confusion matrix of *AUC* values corresponding to the decision region shown in Figure 5.7. The value $M = 0.96$. The asterisk indicates *AUC* values that appear to result from an ideal classifier, but in fact resulted from rounding to two decimal places.

| | **Humpback** | **Right** | **Minke** |
|---|---|---|---|
| **Bowhead** | 0.94 | 1.00* | 0.92 |
| **Humpback** | | 1.00 | 0.96 |
| **Right** | | | 0.96 |

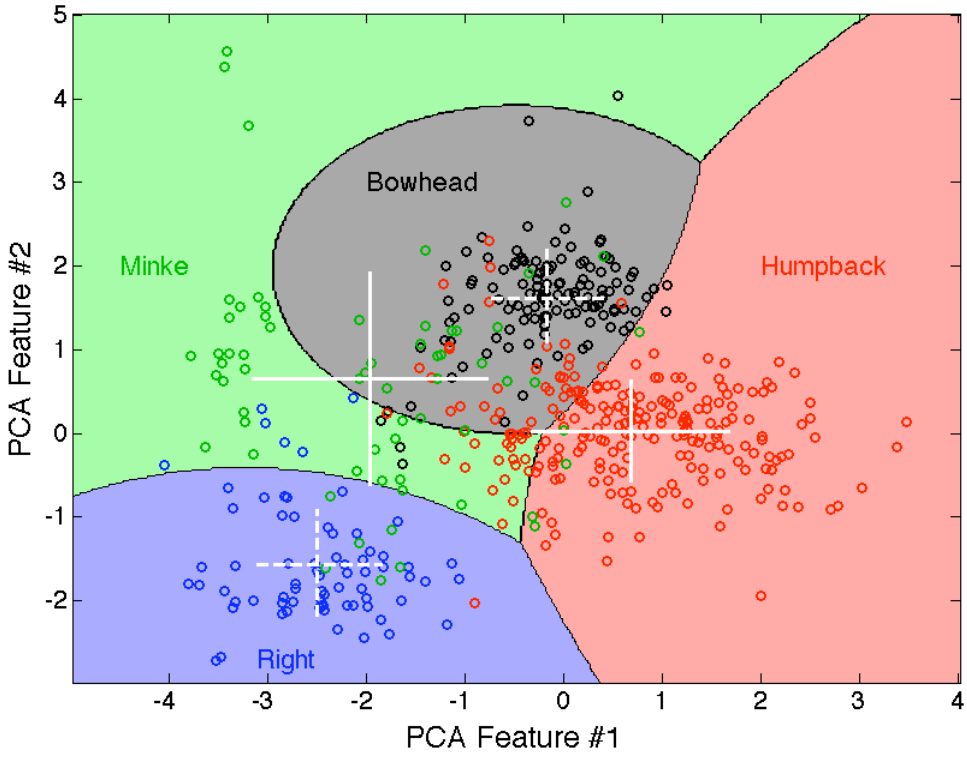


Figure 5.7 Decision region for multiclass classification of baleen whale vocalizations. Classification was performed with eight selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

Classification over all classes improved slightly ($\Delta M = 0.02$) when 8 features were selected compared to 25 selected features. There was a measurable increase in the separation of the bowhead/humpback classes ($\Delta AUC = 0.14$) and bowhead/right whale classes ($\Delta AUC = 0.12$); however discrimination of the minke whale vocalizations deteriorated. With 25 features selected the minke whale class had little overlap with any other class; however, with 8 features selected there was overlap with all the other classes.

Table 5.7 lists the top three features with the highest weighting for classification with each of 25 and 8 selected features and Figure 5.8 plots the feature weighting in the first two principal components. None of the same features were ranked in the top three highest weighted features for either number of selected features. As was the case when all whale species were classified (Section 5.1.1), the zero feature weightings corresponding to features that were not one of the 8 selected features, do not necessarily

coincide with features with the lowest weighting in the principal components containing 25 features. The relative weighting of features in each case are significantly different, reflecting the differences in the methods for selecting and weighting features.

Table 5.7 Three highest weighted features using 25 and 8 selected features for baleen whale classification.

| **25 Selected Features** | **8 Selected Features** |
|---|---|
| Loudness centroid | Global maximum sub-band attack time |
| Psychoacoustic bin-to-bin difference | Psychoacoustic maxima-to-spectral-bins ratio |
| Integrated loudness | Pre-attack psychoacoustic maxima-to-spectral-bins ratio |

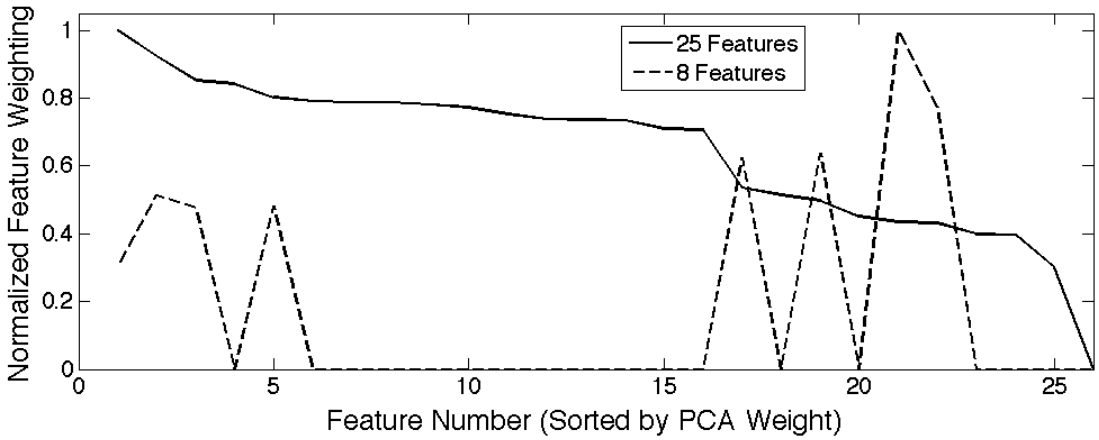


Figure 5.8 Normalized weighting of features in the first two principal components. Features are sorted from largest PCA feature weighting to smallest based on PCA with 25 selected features. These eigenvectors correspond to the decision regions shown in Figure 5.6 and Figure 5.7. Peaks are connected merely for visualization purposes and are not intended to imply that the data are continuous.

## 5.2 BINARY CLASSIFICATION

Binary classification provides insight into the aural features that are most important for distinguishing between two different baleen whale species. An improvement in classification results can be expected when performing binary classification because features are chosen and weighted based on how well they distinguish between the two classes considered. In multiclass classification there may be patterns within the dataset other than species-specific patterns; for example, features may be selected to capture the

significant differences between the clicks in the sperm whale class and moan-like sounds produced by each of the four baleen whale species, rather than the more subtle differences between each of the five classes.

### 5.2.1 Bowhead and Humpback

Bowhead and humpback whales provide an interesting case for binary classification. These two species were included in the dataset primarily because of the similar characteristics of their vocalizations – the frequency content and duration of the vocalizations are similar enough that many types of automatic detection/classification algorithms produce inaccurate results. Similarities between these vocalizations were also noted during the multiclass classification results presented previously (Sections 5.1.1 and 5.1.2) – there was a significant amount of overlap between these classes in the decision regions and the corresponding pairwise *AUC* values were found to be the lowest pairwise *AUC* values (except for the right/minke pair when $c = 5$ and 20 features were selected).

The plot of cumulative Fisher score (Figure 5.9) was examined to determine the number of features to include in the principal components. There is a bend in the plot at 5 features and the curve begins to flatten out at 20 features. Thus, classification was performed with both 20 and 5 features selected.

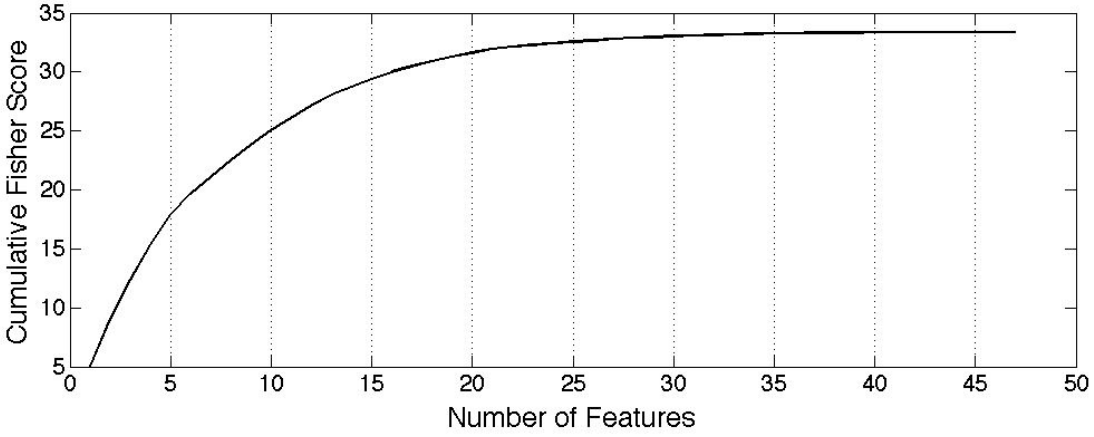


Figure 5.9 Cumulative Fisher score with respect to number of features for bowhead and humpback whales. Note that the features have been sorted so that the first feature has the largest Fisher score. Points are connected for visualization purposes and not intended to imply the data are continuous.

The decision region corresponding to classification with 20 selected features is shown in Figure 5.10 and the ROC curve is plotted in Figure 5.11. The classification accuracy was 88%. There is more variation in the humpback class than the bowhead class,

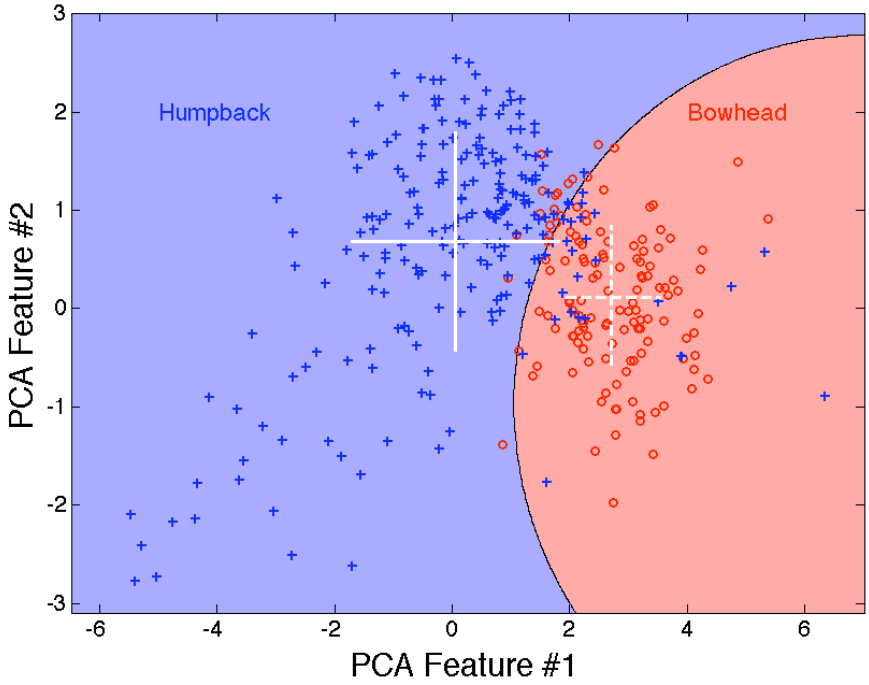


Figure 5.10 Decision region for binary classification of bowhead and humpback vocalizations. Classification was performed with 20 selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

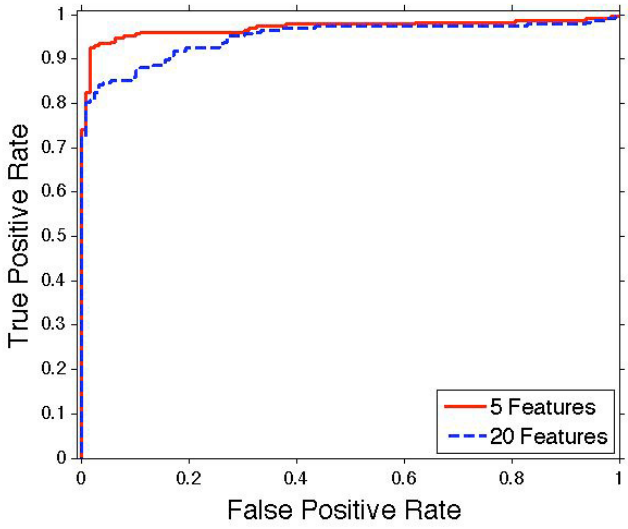


Figure 5.11 Bowhead and humpback ROC curves for classification with 5 and 20 selected features, corresponding to decision regions in Figure 5.10 and Figure 5.12.

corresponding to the larger variation in the sound of humpback whale vocalizations and types of units included in the dataset. There is a relatively small amount of overlap between the two classes – the *AUC* was 0.95 and the equal error rate was 12%. Almost 60% of the variance in the 20 selected features was maintained in the first two principal components ($p_2 = 0.58$).

Classification was also performed with only five features selected. The corresponding ROC curve and decision region are shown in Figure 5.11 and Figure 5.12, respectively. Once again there is more variance evident in the humpback class than in the bowhead class. The value $p_2 = 0.87$ indicates that most of the variance contained in the five selected features was represented by the first two principal components. The classification accuracy was 89%, the *AUC* was 0.97 and the equal error rate was 6%. These values all represent improved classification performance when 5 features are selected instead of 20.

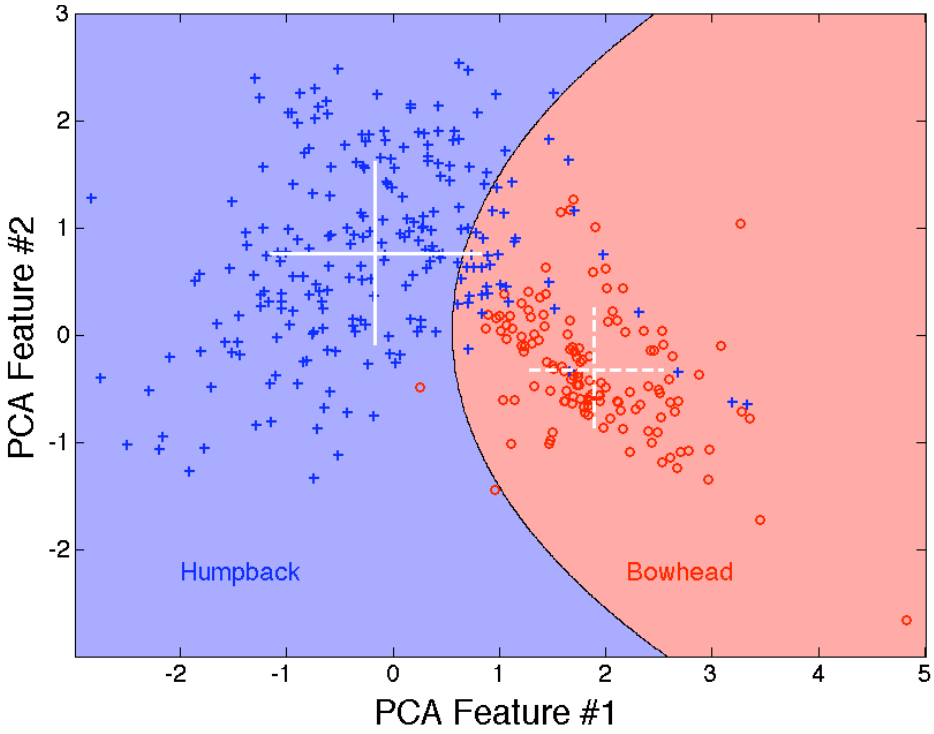


Figure 5.12 Decision region for binary classification of bowhead and humpback vocalizations. Classification was performed with five selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

The three features with the largest weightings in the principal components, for the cases of each of 20 and 5 selected features, are listed in Table 5.8. No features appear in both sets; in fact, the features with large weightings are remarkably different between the two cases. When 20 features were selected time-frequency features were highly ranked, whereas when 5 features were selected purely spectral features had more importance. The highest weighted features when 20 features were selected did not necessarily correspond to the 5 selected features. The 5 features had little similarity in weightings within the principal components compared to the 20 selected features.

Table 5.8 Three highest weighted features using 20 and five selected features for bowhead and humpback classification.

| **20 Selected Features** | **5 Selected Features** |
|---|---|
| Local mean sub-band decay slope | Integrated loudness |
| Global mean sub-band attack slope | Peak loudness value |
| Global mean sub-band decay slope | Local maximum sub-band attack time |

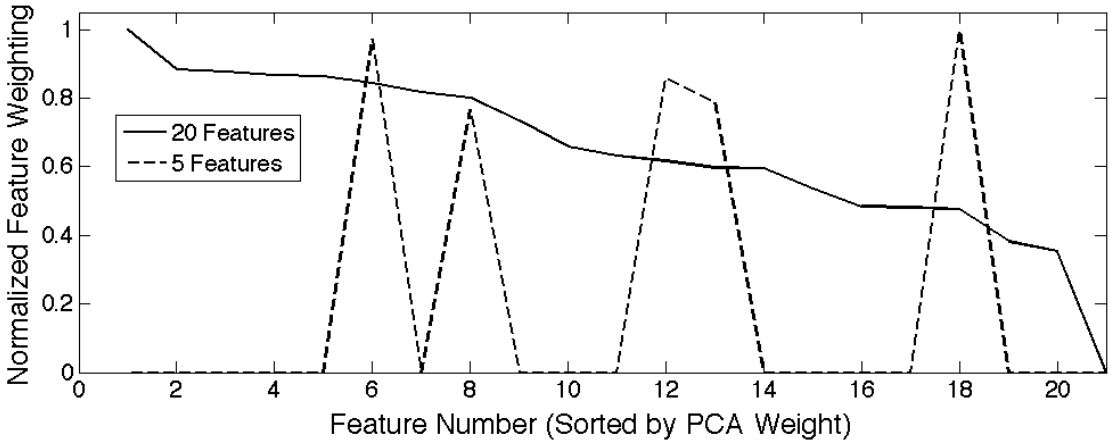


Figure 5.13 Normalized weighting of features in the first two principal components. Features are sorted from largest PCA feature weighting to smallest based on PCA with 20 selected features. These eigenvectors correspond to the decision regions shown in Figure 5.10 and Figure 5.12. Peaks are connected merely for visualization purposes and are not intended to imply that the data are continuous.

### 5.2.2 Summary of Baleen Whale Binary Classification Results

Binary classification of baleen whale vocalizations was performed with all six possible pair combinations. Classification was performed with five selected features, since it has been found in previous sections that a smaller number of selected features consistently produced better classification results. Decision regions for each classification are depicted in Figure 5.14. The greatest overlap of classes occurred for the bowhead/ humpback classification. All other classifications resulted in only a small amount of overlap between classes – there were few misclassifications among the binary decision regions (excluding bowhead/humpback) so that classification results are near-ideal. The humpback and minke classes displayed the most within-class variance in all classifications in which they were included. It is not surprising that the humpback class displayed relatively large within-class variance since four distinct units were included – each of which had complex aural characteristics. The large within-class variance for minke whales may result from sound propagation effects; some of the minke vocalizations in the dataset exhibited high frequency overtones with energy similar to the fundamental frequency, whereas other minke vocalizations either had weak high frequency overtones or no high frequency overtones at all. The energy in the high frequency overtones may have been reduced due to propagation through the water.

The ROC curves (Figure 5.15) corresponding to the decision regions shown in Figure 5.14 confirm that classification was near ideal in all six baleen whale binary classification cases. The lowest *AUC* value (see Table 5.9) was 0.97 for the bowhead/humpback classification; the majority of *AUC* values were 1.00. The largest equal error rate was 7% corresponding to the minke/right whale classification. These binary classification cases provide a real example for the importance of reporting both the *AUC* value and equal error rate for binary classifications; the smallest *AUC* value does not have to correspond to the largest equal error rate. When the smallest *AUC* value and largest equal error rate do not belong to the same ROC curve, it indicates that the corresponding ROC curves cross each other at least once. ROC curves that cross indicate that one classifier does not perform the best for all threshold values. There are only a few threshold values for which the bowhead/humpback classification performs better than the right/minke classification

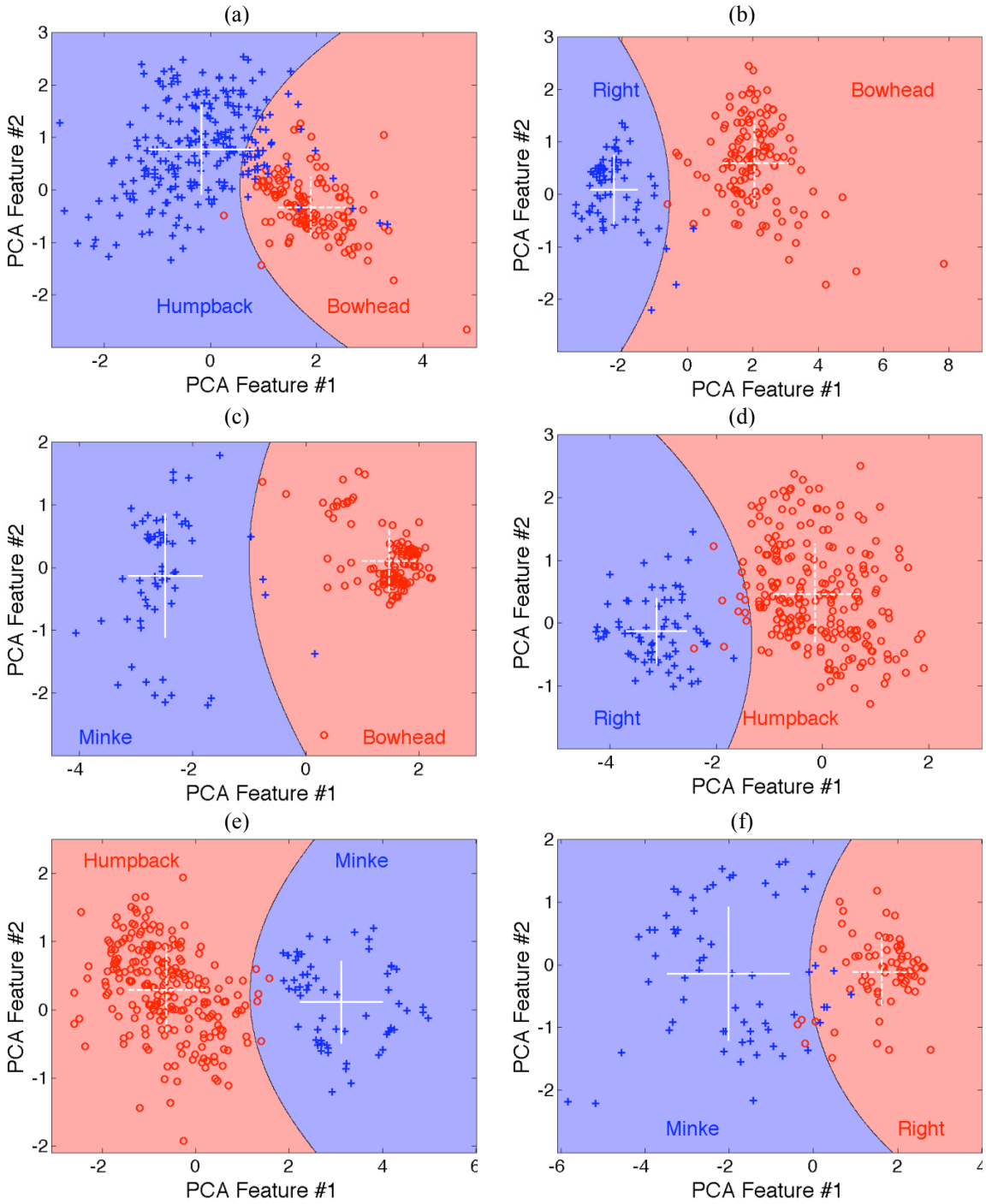


Figure 5.14 Binary decision regions for baleen whale classifications involving (a) bowhead and humpback, (b) bowhead and right whale, (c) bowhead and minke, (d) humpback and right whale, (e) humpback and minke, and (f) minke and right whale. Five features were selected to include in the principal components – note that a different classifier model (i.e. different features were selected and combined in the principal components) was used to generate each of the decision regions.

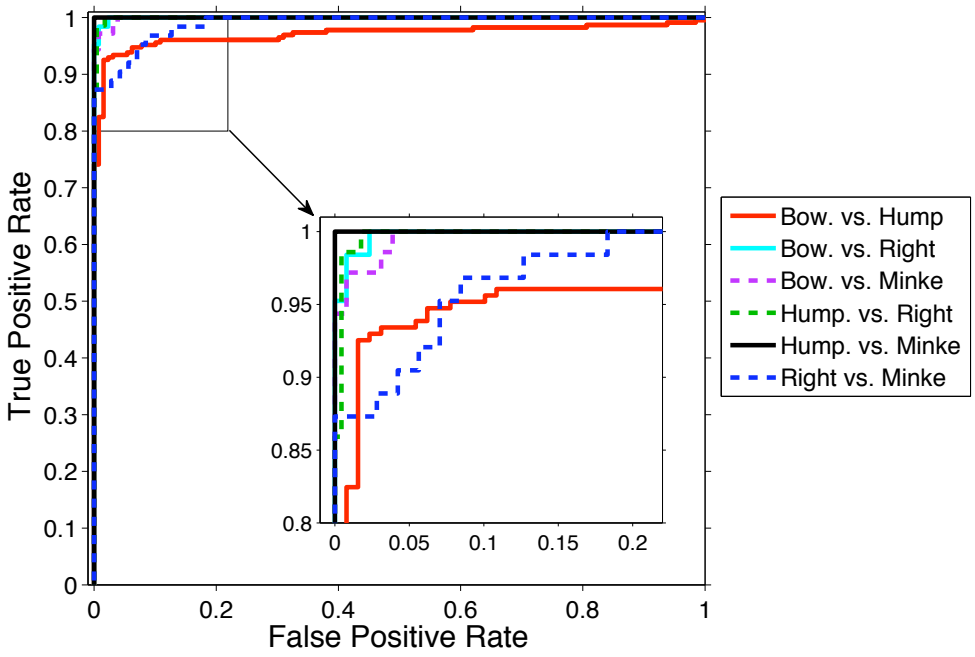


Figure 5.15 Binary ROCs using five selected features and two principal components corresponding to the decision regions in Figure 5.14. The inset shows a zoomed in view of the ROC curves.

Table 5.9 The accuracies, *AUC* values, and equal error rates for binary classification of baleen whales using five selected features. These values correspond to the ROC curves in Figure 5.15. The asterisk indicates *AUC* values that appear to result from an ideal classifier, but in fact resulted from rounding to two decimal places.

| **Whales** | **Accuracy** | ***AUC*** | **Equal error** |
|---|---|---|---|
| Bowhead vs. humpback | 89% | 0.97 | 6% |
| Bowhead vs. right | 97% | 1.00* | 3% |
| Bowhead vs. minke | 98% | 1.00* | 2% |
| Humpback vs. right | 97% | 1.00* | 1% |
| Humpback vs. minke | 98% | 1.00 | 0% |
| Right vs. minke | 93% | 0.99 | 7% |

as can be seen in Figure 5.15. Classification of bowhead and humpback vocalizations was the least accurate (89%) and classifications of bowhead/minke and humpback/minke (both 98%) were the most accurate. Performance results reflect the aural similarities of the vocalizations – bowhead and humpback vocalizations sound more similar to each other than bowhead and minke or humpback and minke vocalizations.

The features with the three largest weight values in the principal components for each of the baleen whale binary classification pairs are listed in Table 5.10. Some of the same features are chosen multiple times, e.g. local maximum sub-band attack time, peak loudness value, and mean sub-band correlation each received high weighting in two of the six baleen whale binary classifications. Features that were chosen multiple times can be recognized as important for between-class discrimination of baleen whale vocalizations.

Table 5.10 Three highest weighted features using five selected features for each pair of baleen whales, shown in descending importance from left to right. Features represented in italics were important for at least two of the binary classification pairs.

| | | | |
|---|---|---|---|
| **Bowhead/ Humpback** | Integrated loudness | *Peak loudness value* | *Local maximum sub-band attack time* |
| **Bowhead/ Right** | Pre-attack psychoacoustic maxima-to-spectral-bins ratio | *Local maximum sub-band attack time* | Psychoacoustic maxima-to-spectral-bins ratio |
| **Bowhead/ Minke** | Frequency of maximum sub-band correlation | Frequency of local minimum sub-band attack slope | Frequency of global minimum sub-band attack slope |
| **Humpback/ Right** | *Mean sub-band correlation* | Pre-attack integrated loudness | *Peak loudness value* |
| **Humpback/ Minke** | *Mean sub-band correlation* | Frequency of local maximum sub-band attack slope | Frequency of local minimum sub-band decay slope |
| **Right/ Minke** | Pre-attack loudness centroid | Global maximum sub-band decay time | Local mean sub-band attack time |

### 5.2.3 Sperm Whale Clicks and Baleen Whale Vocalizations

Sperm whale clicks separate out with 100% accuracy in the multiclass plots shown in Figure 5.2b and Figure 5.3, but significant overlap of the baleen whale classes remained. The selection and relative weighting of features in the multiclass case when all five species are included seems to be driven by the significant aural differences between sperm whale clicks and baleen whale vocalizations. This is in evidence partly by comparing the features selected for multiclass classification of all species with 20 features (Section 5.1.1) and baleen whale species with 8 features (Section 5.1.2) –

relative feature weightings are noticeably different for both classification cases. To confirm that the selection of features in the multiclass case is dominated by the difference between sperm whale clicks and baleen whale vocalizations, binary classification was performed where one class was composed of sperm whale clicks and the second class comprised vocalizations from all four baleen whale species. The features used for this binary classification were the top five ranked features (i.e. highest weighted features in the principal components) for classification of all five cetacean species. Five features were used for classification so as to produce a relatively simple transformation between the feature space and PCA space.

Results of aural classification of baleen whale vocalizations and sperm whale clicks are presented as the decision region in Figure 5.16. The decision region reveals a clear separation of sperm whale clicks and baleen whale vocalizations. Even though there are vocalizations from four species of baleen whale – representing a large variety of call types – data points belonging to the baleen whale class are clustered together.

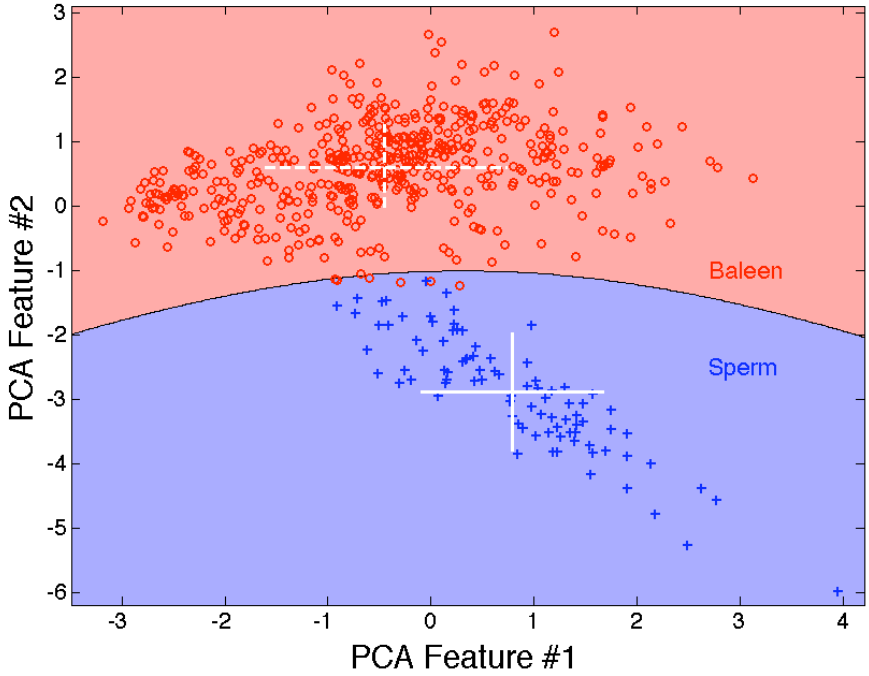


Figure 5.16 Decision region for binary classification of sperm whale clicks and baleen vocalizations. Classification was performed with the five most important features for multiclass classification of all cetacean species. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

Classification accuracy was 98% – only a few baleen whale vocalizations were misclassified and all sperm whale clicks were correctly classified. The *AUC* was 1.00 (rounded up) and equal error rate was between 0% and 1%. The aural classifier had no difficulty distinguishing between these aurally distinct sounds using the five most important features for multiclass classification of all cetacean species. This supports the hypothesis that the selection and weighting of features in the multiclass case is likely dominated by the obvious aural difference between sperm whale clicks and baleen whale vocalizations.

## 5.3 DISCUSSION

### 5.3.1 Number of Features to Select

One might assume that including more features for classification would enhance performance results by providing more information about the patterns between classes; however, this was not the case in the preceding sections. For multiclass classification (Section 5.1) the *M*-measure remained the same for classification of all species with either 30 or 20 features selected. Perhaps more surprising, the *M* value was *larger* when baleen whales were classified using 8 selected features compared to 25 selected features; and the *AUC* was *larger* and equal error rate *smaller* when 5 features were used for classification of bowhead and humpback vocalizations instead of 20 features. In all cases, the amount of relative variance contained in the first two principal components, as measured by $p_2$, increased when fewer features were used for PCA.

Since no theoretical model or solution is available, it is difficult to determine the optimal number of features to select for best classifier performance without performing classification with all possible numbers of selected features. This presents a paradox – to select the optimal number of features for classification, the classifier performance with respect to number of selected features must first be known. The *M* value was plotted (Figure 5.17) with respect to number of selected features (starting with two features and ending with all 44 non-redundant features for multiclass classification of bowhead, humpback, right, minke and sperm whales. The maximum value of *M* (0.99) was obtained when only 5 features were selected. This plot reveals that increasing the

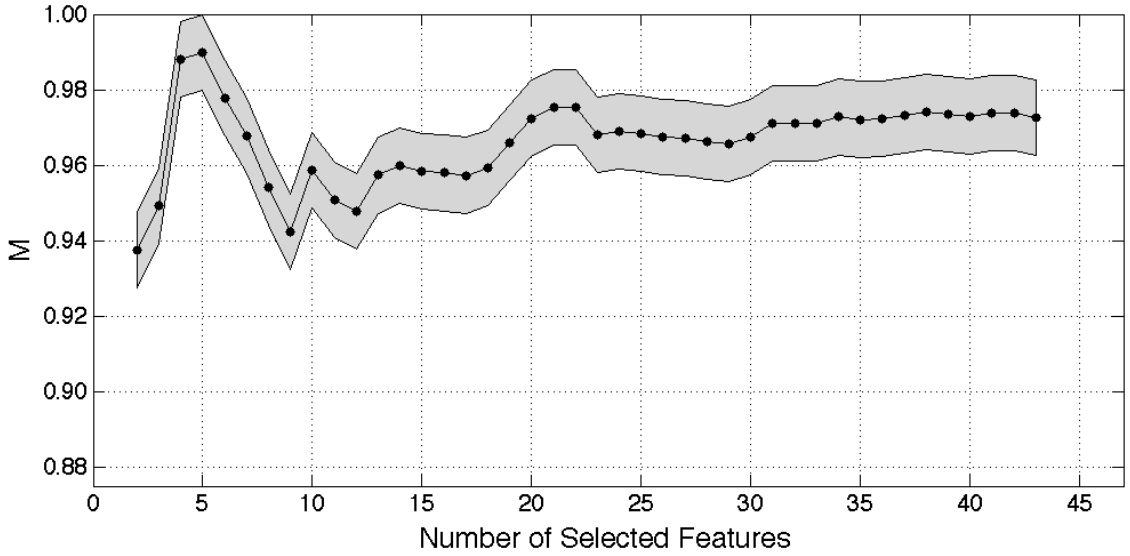


Figure 5.17 Performance results for classification of all cetacean species with respect to number of features included in the principal components. The grey region represents the estimated error resulting from calculation of the *M*-measure. Points are connected merely for visualization purposes and are not intended to imply that the data are continuous.

number of selected features does not necessarily correspond to increased performance. With the exception of the large peak at five selected features, there appears to be a general increasing trend in *M* that begins to level off around 31 selected features.

For completeness, the multiclass all-species decision region generated using five selected features is shown in Figure 5.18. The *M* value was 0.99 indicating near-ideal performance. The classification accuracy was increased significantly to 89%. There was little overlap between any of the classes as confirmed by the large pairwise *AUC* values in Table 5.11 – in fact even the bowhead/humpback pair, which showed a lot of overlap in all previous multiclass classification results, separated out well. The bowhead/ humpback pairwise *AUC* value was as high as for binary classification of bowhead and humpbacks with five selected features. This was the first multiclass classification result that had misclassifications including sperm whales – there was one sperm whale click misclassified as a right whale vocalization and two right whale vocalizations misclassified as sperm whale clicks. Both misclassified right whale vocalizations were gunshot sounds (see Figure 3.5c), which have short duration and energy spread equally

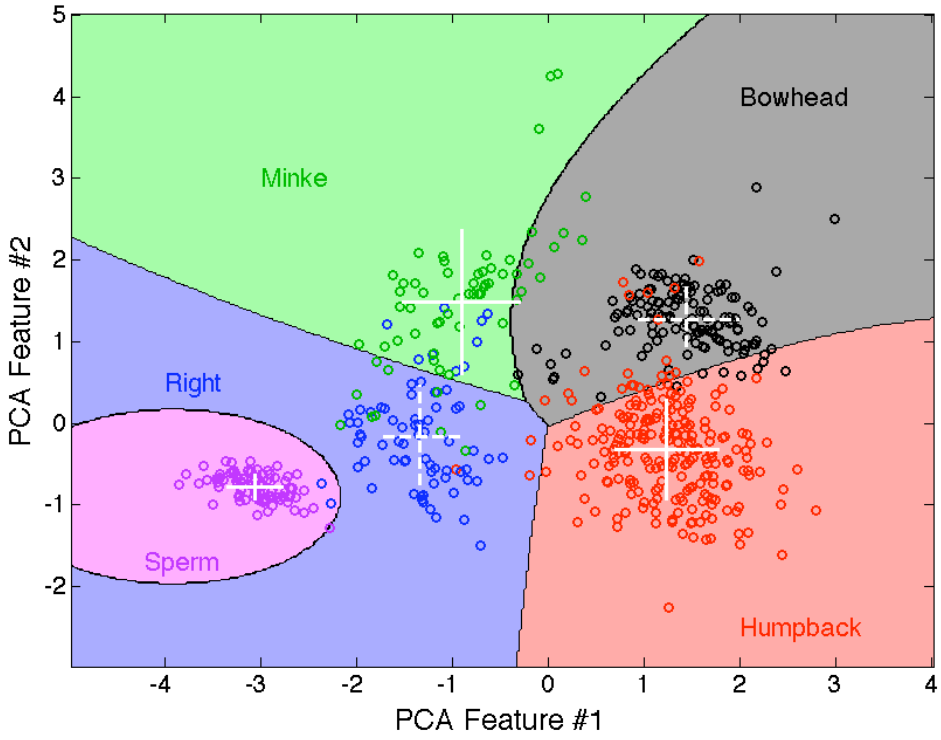


Figure 5.18 Decision region for multiclass classification of all cetacean vocalizations. Classification was performed with five selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

Table 5.11 Confusion matrix of *AUC* values corresponding to the decision region shown in Figure 5.18. The value $M = 0.99$. The asterisk indicates *AUC* values that appear to result from an ideal classifier, but in fact resulted from rounding to two decimal places.

| | **Humpback** | **Right** | **Minke** | **Sperm** |
|---|---|---|---|---|
| **Bowhead** | 0.97 | 1.00* | 1.00* | 1.00 |
| **Humpback** | | 1.00* | 1.00* | 1.00 |
| **Right** | | | 0.94 | 1.00* |
| **Minke** | | | | 1.00 |

across the frequency band of the vocalization, similar to sperm whale clicks. Also of note is the spread of sperm whale data points; the within-class covariance and correlation are -0.02 and -0.47 respectively. In this case, the assumption is valid that the off-diagonal components of the covariance matrix are zero, unlike for the previous classification results.

Table 5.12 Five highest weighted features for multiclass decision region with 20 selected features shown in Figure 5.3, and multiclass decision region with 5 selected features shown in Figure 5.18

| **Multiclass, 20 Features** | **Multiclass, 5 Features** |
|---|---|
| Integrated loudness | Duration |
| Psychoacoustic bin-to-bin difference | Mean sub-band correlation |
| Pre-attack peak loudness value | Pre-attack psychoacoustic maxima-to-spectral-bins ratio |
| Pre-attack psychoacoustic maxima-to-spectral-bins ratio | Psychoacoustic maxima-to-spectral-bins ratio |
| Mean sub-band correlation | Loudness centroid |

Table 5.12 contains the five highest weighted features for classification of the five cetacean species for the multiclass classification, with 20 and 5 features selected. The only feature in common between the multiclass 5-feature and 20-feature classification cases was pre-attack psychoacoustic maxima-to-spectral bins ratio; otherwise, different features were selected and highly ranked. It is clear that in the five-feature case the features were selected and weighted to discriminate between all five classes rather than just to discriminate between baleen and sperm whales.

The *AUC* results with respect to number of selected features for the binary classification of bowhead and humpback vocalizations is shown in Figure 5.19. The maximum *AUC* value of 0.97 occurred when three features were selected. The *AUC* result for three selected features is not significantly different than for five selected features. The equal error rate when three features are selected is slightly lower (5% for three features and 6% for five features). With the exception of a few local minima, there appears to be a general decreasing linear trend in the *AUC* values with respect to number of selected features – this is the opposite of the trend that was noted in the plot of *M* versus number of features selected (Figure 5.17).

The decision region for classification of bowhead and humpback vocalizations with three selected features is shown in Figure 5.20. The decision region shows that classification with three selected features minimally improved results – a few more humpback vocalizations (10 more out of 228 humpback vocalizations in the test set) were correctly classified and all 129 bowhead vocalizations were correctly classified, whereas two

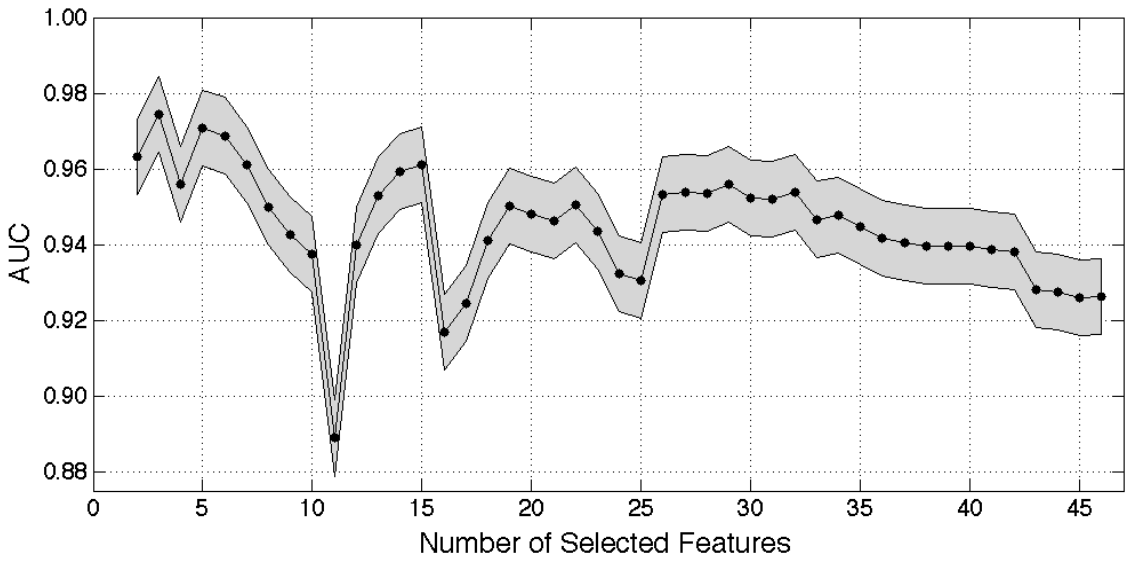


Figure 5.19 Performance results for classification of bowhead and humpback vocalizations with respect to number of features included in the principal components. The grey region represents the estimated error resulting from calculation of the *AUC*. Connected points are merely for visualization purposes and are not intended to imply that the data are continuous.

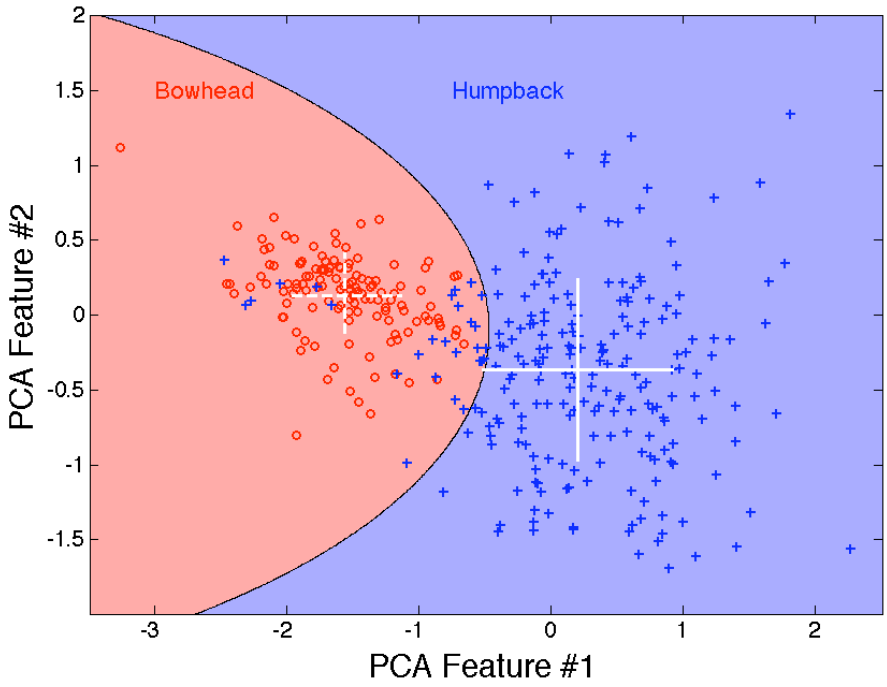


Figure 5.20 Decision region for binary classification of bowhead and humpback vocalizations. Classification was performed with three selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

Table 5.13 Features selected for binary classification of bowhead and humpback whales using either three or five selected features. Features are listed in order of highest weighting in principal components to lowest

| **3 Features** | **5 Features** |
|---|---|
| Peak loudness value | Integrated loudness |
| Mean sub-band correlation | Peak loudness value |
| Global maximum sub-band decay time | Local maximum sub-band attack time |
| | Global maximum sub-band decay time |
| | Mean sub-band correlation |

bowhead vocalizations had been misclassified when five features were selected. The classification accuracy increased slightly to 92% and the equal error rate also improved slightly to 5%. The selected features are listed in Table 5.13 by order of PCA feature weighting. All three of the features selected for three-feature classification were included in five-feature classification; however, the three features did not correspond to the three highest ranked features in the five-selected feature model.

The plots of $M$ and $AUC$ with respect to number of selected features confirm that including more features does not necessarily increase classification performance. Thus, a subset of the non-redundant features should be selected prior to performing PCA. Since the Fisher Linear Discriminant score was used to rank features that best separated the classes, it seemed logical to use the cumulative Fisher score to select the number of features to be projected onto the 2D PCA space; however, this method did not produce the best performance results. The discrepancy is likely due to the different goals of the Fisher score and PCA. The Fisher score ranked features according to their discriminability, whereas PCA placed more weight on features that maintain the variance over the whole dataset. This difference in emphasis was noted because features with zero weighting did not necessarily coincide with the lowest weighted features in the principal components when different numbers of features were selected (see Figure 5.4, Figure 5.8 and Figure 5.13). Therefore, it is difficult to predict the optimal number of features to select.

It seems intuitive to assume that including all features that provide information useful in separating the class means (i.e. relatively large values of the Fisher score) should enhance

classifier performance; however, the results presented here do not support this. Duda *et al*. [24] suggest that this may be because of an incorrect model (e.g. the assumption that each class is Gaussian distributed in the PCA space) or that there is an insufficient number of training samples to accurately estimate the class distributions. In this case, it is assumed that a limited number of features should be used for classification because the dataset is too small to provide sufficient information to accurately estimate the presumed Gaussian likelihood distributions for each class.

### 5.3.2 Linear Trend within Sperm Whale Class in Multiclass Decision Regions

The multiclass decision regions that included all five species of cetaceans in the dataset had a notable feature – the arrangement of sperm whale clicks. When either 30 or 20 features were selected the sperm whale clicks were arranged in an oblong pattern that was characterized by non-zero covariance of the principal components. However, when only five features were selected for multiclass classification the placement of sperm whale data points in the PCA space was as expected – a tight cluster of points that had near-zero covariance of the principal components. This trend in the sperm whale click data points did not occur within any of the other whale classes.

Including too many features in the principal components probably over-described the contrasts between sperm whale clicks and baleen whale vocalizations. The binary classification results of sperm whale clicks and baleen whale vocalizations presented in section 5.2.3 indicated that the choice and relative importance of the 20 and 30 selected features in the multiclass decision regions was likely dominated by the aural distinctness of the sperm whale clicks and baleen whale vocalizations; however, most of the five features selected for the decision region shown in Figure 5.18 were different than the highly ranked features in the principal components for 20-feature and 30-feature classification. These five features described the relationship between all species' vocalizations rather than just the obvious differences between sperm whale and baleen whale vocalizations. Another linear arrangement of sperm whale click data points in the PCA space will be observed and discussed in Sections 6.2.1 and 6.2.2.

The linear trend in the sperm whale class is further investigated in Section 6.2.2 when an example will be presented that allows a relatively simple analysis of the linear trend observed in the PCA space; the conclusions from that analysis will be tested here. Histograms of the integrated loudness (Figure 5.21) and psychoacoustic bin-to-bin difference (Figure 5.22) values, which were the two highest weighted features in the multiclass classification scenario, were generated to determine the amount of within-class variance for each feature. All baleen whales are treated as a single class and sperm whales as another class (as with binary classification). It should be noted that each feature is unitless due to the method in which features were normalized (as described in Section 2.1.4). The histograms exhibit overlap of the classes for each feature; however when the two features are considered together (as in Figure 5.23) without performing PCA, the relationship between the two features becomes clear. There is an emerging within-class linear arrangement of the data points evident when the top two features are plotted. The classes both form an oblong shape with large variance along the semi-major axis of the scatter and relatively little variance along the semi-minor scatter axis. It is likely this relationship between features that is the cause of the linear arrangement of sperm whale data points in the multiclass decision region. The other features included in

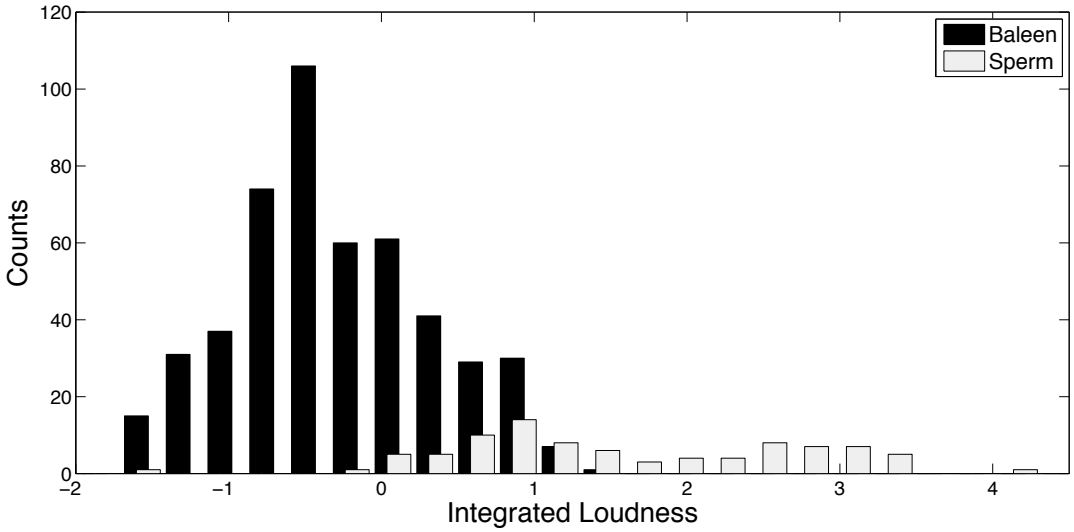


Figure 5.21 Histogram of integrated loudness values, the highest ranked feature in multiclass classification of all five cetacean species.

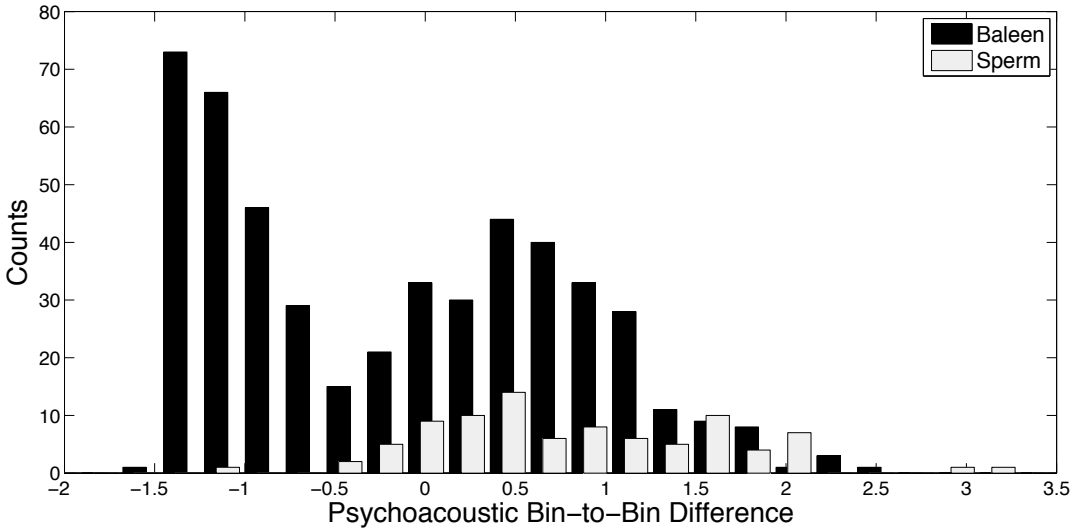


Figure 5.22 Histogram of psychoacoustic bin-to-bin difference values, the second highest ranked feature in multiclass classification of all five cetacean species.

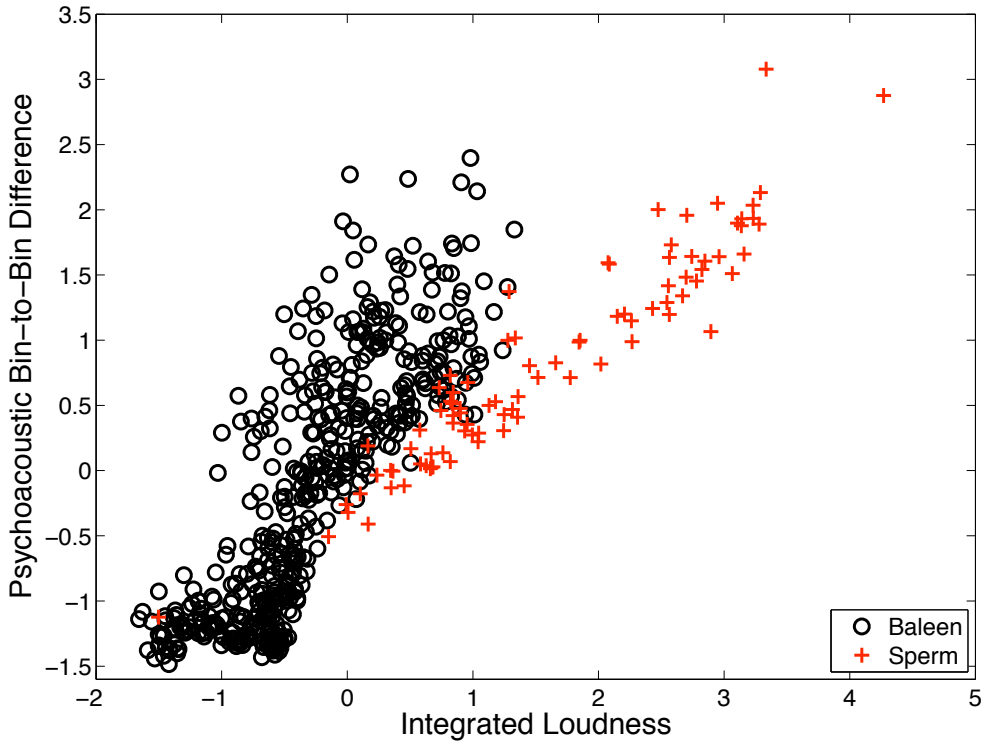


Figure 5.23 Two highest ranked features in the principal components for the multiclass decision regions shown in Figure 5.2b and Figure 5.3. These results were plotted without performing PCA on the displayed features.

the principal components for multiclass classification maintain (or possibly add to) this relationship within the sperm whale class, but allow for additional between-class discrimination of the baleen whales by increasing the scatter in all directions.

### 5.3.3 Important Aural Classification Features

An important goal of this research was to identify the aural features most useful for classification of cetacean vocalizations. The six features that were most often selected and received large weight values in the first two principal components were mean sub-band correlation, pre-attack psychoacoustic maxima-to-spectral-bins ratio, integrated loudness, psychoacoustic maxima-to-spectral-bins ratio, peak loudness value, and psychoacoustic bin-to-bin difference. Table 5.14 lists all the features that were selected and the number of times they were included in the previous tables (Table 5.4, Table 5.7, Table 5.8, Table 5.10, Table 5.12, and Table 5.13) of features that were highly ranked in the principal components. In general, purely spectral features were selected more often than time-frequency features, possibly indicating that the vocalizations of each whale species have distinct spectral characteristics.

Table 5.14 Number of times each feature was included in features with highest weight value in the principal components in the 13 different classification cases discussed. The tally included features in Table 5.4, Table 5.7, Table 5.8, Table 5.10, Table 5.12, and Table 5.13.

| **Feature** | **Number of Occurrences (of a possible 13)** |
|---|---|
| Mean sub-band correlation | 5 |
| Pre-attack psychoacoustic maxima-to-spectral-bins ratio | 5 |
| Integrated loudness | 4 |
| Psychoacoustic maxima-to-spectral-bins ratio | 3 |
| Peak loudness value | 3 |
| Psychoacoustic bin-to-bin difference | 3 |
| Loudness centroid | 2 |
| Global maximum sub-band decay time | 2 |
| Local maximum sub-band attack time | 2 |
| Duration | 1 |
| Frequency of global minimum sub-band attack slope | 1 |

| Feature | Number of Occurrences (of a possible 13) |
|---|---|
| Frequency of local maximum sub-band attack slope | 1 |
| Frequency of local minimum sub-band attack slope | 1 |
| Frequency of local minimum sub-band decay slope | 1 |
| Frequency of maximum sub-band correlation | 1 |
| Global mean sub-band attack slope | 1 |
| Global mean sub-band decay slope | 1 |
| Global maximum sub-band attack time | 1 |
| Local mean sub-band decay slope | 1 |
| Local mean sub-band attack time | 1 |
| Pre-attack integrated loudness | 1 |
| Pre-attack loudness centroid | 1 |
| Pre-attack peak loudness value | 1 |

Young [20] discusses the origin of all the perceptual features and how they relate to the aural characteristics of a particular sound – his work was referenced to describe the six important perceptual features. The mean sub-band correlation is the only time-frequency feature that was frequently identified as an important feature for inter-species cetacean classification. This feature describes the synchronicity of the signal across all harmonics, i.e. it determines if harmonics rise and fall at the same rate and time or if harmonics are independent of each other. This feature is quantitatively determined by correlating the filter bank channels – large correlation indicates that the harmonics are synchronous. Humpback vocalizations are highly correlated as is evident from the spectrograms of humpback units presented in Figure 3.4; the humpback units tend to have several obvious overtones that appear identical to each other except for a frequency shift.

Both the peak loudness value and integrated loudness features are easily determined from the perceptual loudness spectrum; the peak loudness value is the maximum value of the perceptual loudness spectrum and the integrated loudness is simply defined as the area under the perceptual loudness spectrum. Since the perceptual loudness spectrum is basically a psychoacoustic power spectrum, the peak loudness value and the integrated loudness are analogous to the maximum power and the total power present in the signal, respectively. Generally, humpback vocalizations had the highest peak loudness values

and right whales the lowest; sperm whale clicks had the largest integrated loudness values and minke whales the lowest. Large peak loudness values do not necessarily correspond to the largest integrated loudness values because integrated loudness is determined across all frequencies, but peak loudness value corresponds to a specific frequency. For example, humpback whales have the highest peak loudness values; however, sperm whales have larger values of integrated loudness because the energy of sperm whale clicks is spread evenly across the frequency range, whereas the energy in humpback units is condensed into a relatively narrow frequency band.

The psychoacoustic maxima-to-spectral-bins ratio (PMSBR) is used to describe the roughness of the loudness spectrum. The presence/absence of spectral peaks has been identified as an important property for aural discrimination. This is represented as an aural feature as the ratio of the number of local maxima to the total number of spectral bins in the perceived loudness spectrum (number of channels in the filter bank). Generally, it was found that the perceptual loudness spectrum of sperm whale clicks was rougher than that of baleen whale vocalizations. The pre-attack PMSBR is also an important feature. The difference between these two features is that the pre-attack PMSBR is determined from the pre-attack component of the signal rather than from the whole signal as for PMSBR. The musical acoustics literature identified the segment of the signal prior to the most significant attack as an important discrimination cue that can be qualitatively described as the presence or absence of high frequency inharmonic noise. Thus, features extracted from the pre-attack signal component are intended to identify the presence/absence of inharmonic noise. Identification of the pre-attack component of the signal is discussed in Appendix A.

The psychoacoustic bin-to-bin difference (BBD) is also a measure of the roughness of the perceptual loudness spectrum. This aural feature describes the BBD across the loudness function for each vocalization. The differences in the specific loudness function between pairs of adjacent frequency bins are calculated and the results are averaged for all pairs to determine the psychoacoustic BBD. Generally, it was found that bowhead and right

whale vocalizations had small psychoacoustic BBD values, whereas humpback and sperm whales had larger values.

### 5.3.4 Comparison of Aural Classification Results with Literature Results

Much of the research on marine mammal detection and classification presented in the literature has a different goal than the aural classification task presented in this thesis. Detection/classification is often used to perform population surveys or behavioural studies that focus on a single species, so most often detection and classification are inseparable. In other words, the methods used to detect a vocalization are specifically tuned to a certain species so that upon detection there is little doubt as to the species that produced the vocalization. For example, Mellinger and Clark [46] used spectrogram-correlation and a simple matched filter to compare results of detection/classification of bowhead sounds from MobySound (i.e. using the same bowhead dataset as employed by this thesis). Both the spectrogram-correlation and matched filter methods require a template of a typical bowhead sound with which to correlate the input signal. By detecting bowhead sounds in the dataset, Mellinger and Clark were, in essence, classifying the bowhead sounds against all other sounds in the dataset – primarily vocalizations of bearded seals. They were able to detect bowhead vocalizations with accuracies of 84.2% using a matched filter and 99.1% using spectrogram correlation. Their results are highly accurate; however they do not present results for the more challenging case of bowhead and humpback classification. Their method is specifically tuned to recognize a single species, whereas the aural classifier is capable of simultaneously classifying several different species.

North Atlantic right whales have been the focus of many automatic detection studies using passive acoustics, including population estimates and ship avoidance studies. Reports indicate that right whale vocalizations can be confused with the highly vocal humpback whale's vocalizations, many of which are similar in frequency content and duration to right whale sounds [15]. The risk associated with humpback whale presence is usually lower than that of right whale presence because of the relatively large number of humpback whales (population is not under stress) compared to the endangered status

of North Atlantic right whales. Gillespie developed a right whale detector that was able to detect right whale sounds with ~60% efficiency and a false alarm rate of 1 – 2 calls per detector per day, even though tens of thousands of humpback sounds were also present [47]. In another study, Mellinger compared two different right whale up-call detectors, one based on spectrogram correlation and another on neural networks; he found that at a 10% false negative rate (i.e. 10% right whale up-calls are likely to be missed), the neural network method had a false positive rate of 6% and the spectrogram correlation method had a false positive rate of 26% [12]. Baumgartner and Mussoline [48] performed detection and classification on recordings collected in the northwestern Atlantic Ocean that contained sei (a baleen whale), right, and humpback whale vocalizations. Employing features extracted from the spectrogram-track of each detected vocalization, Baumgartner and Mussoline's classifier accurately classified only 52% of the right whale calls.

Compared to the available results of other detection/classification methodologies, the aural classifier performed very well. A significant advantage of the aural classification method is the ability to easily include additional species for classification because it is not specifically tuned for detection/classification of a single species. Given a large enough training set, it is reasonable to assume that the aural classifier would deal with variation in cetacean vocalizations better than correlation techniques.

## 5.4 CONCLUSIONS

The aural classifier was shown to be very effective at discriminating between the cetacean vocalizations in the dataset. Although the dataset was relatively limited, it provided an opportunity to validate the theory that the aural classifier can successfully classify marine mammal vocalizations. Classification performance reflected both the aural similarities and distinctness of the cetacean vocalizations. The classifier was able to classify sperm whale clicks from baleen whale vocalizations with 100% accuracy because these two classes of sounds have distinct aural characteristics, whereas the lowest classifier performance corresponded to the aurally similar bowhead and humpback vocalizations.

The six features found to be the most powerful for discriminating between the cetacean vocalizations in the dataset were mean sub-band correlation, pre-attack PMSBR, integrated loudness, PMSBR, peak loudness value, and psychoacoustic BBD. Since most of these features were obtained from the perceptual loudness spectrum, it was concluded that there are significant differences in the perceived spectral characteristics of these vocalizations.

With the current dataset, it was found that selecting more features did not necessarily result in increased performance – in fact classifier performance deteriorated when large numbers of features were selected. It was suggested that the observed decrease in classifier performance with a larger number of selected features resulted from an insufficient number of samples in the training dataset to accurately describe the class distributions.

High classifier performance was obtained for both the simple binary case and the more complex multiclass case. Including more species for classification in the multiclass case did not result in a significant decrease in performance. The successful multiclass classification results presented in this chapter support the hypothesis that the aural classifier can be trained to simultaneously classify several different marine mammal species.

# CHAPTER 6 SPERM WHALE AND ANTHROPOGENIC TRANSIENT CLASSIFICATION

It was shown in CHAPTER 5 that the aural classifier was able to easily distinguish between sperm whale clicks and baleen whale vocalizations due to obvious differences in aural characteristics; however, many other types of detector/classifier algorithms would be able to recognize the differences in sperm whale clicks and baleen whale vocalizations because of clear differences in their signal characteristics – sperm whale clicks are shorter in duration than any of the baleen whale vocalizations considered and sperm whale clicks are broadband, whereas baleen whale vocalizations are relatively narrowband (i.e. their frequency band is more limited).

A more realistic test case for classification of sperm whale clicks is to compare them to anthropogenic passive transients. Many of the signal properties of sperm whale clicks are similar to a variety of anthropogenic transients. When detecting sperm whale clicks in a signal, both correlation-based (e.g. matched filtering) and energy-based (e.g. band-limited energy detection) methods produce many false detections due to anthropogenic transients. For example, the automatic detector used by Akoostix Inc. to identify sperm whale clicks generated a total of 1495 detections, of which only 178 could be positively identified as sperm whale clicks. Detection of sperm whale clicks is a difficult case and often produces many false alarms because the exponentially averaged detector employed by Akoostix Inc. is triggered by most broadband sounds that have an abrupt start [31]. Sperm whale clicks are more similar in frequency content and duration to anthropogenic transients than to baleen whale vocalizations; therefore, anthropogenic transients may

prove more challenging to classify against sperm whale clicks. Nonetheless, the aural classifier may be used to discriminate between sperm whale clicks and anthropogenic transients because these two classes are aurally dissimilar.

## 6.1 ANTHROPOGENIC TRANSIENT DATASET

Akoostix Inc. compiled a dataset of anthropogenic passive transients for research and development related to the aural classifier [49]. Transients were obtained from the Passive Aural Listening Database (PAL) at the Atlantic Acoustic Data Analysis Centre (ADAC). Transients in this database were previously detected and then aurally classified by expert listeners.

Anthropogenic passive transients were selected from the PAL database based on the following characteristics:

- Sounds could be related to a vessel (e.g. ship, submarine) or other man-made platform (e.g. mooring).
- Events with an abrupt start or stop (e.g. motor start), though only the region in the vicinity of the start or stop is considered
- Duration of transient must be in the range 1 ms – 5 s
- Signal's bandwidth should be greater than the inverse of the duration

Sounds generated by active transducers (e.g. echo sounder), or transients that could be modelled in simple closed-form (e.g. narrow-band pulsed continuous wave), were not considered to be anthropogenic transients for the purposes of this study. Based on these characteristics, six subsets of anthropogenic transients were identified: ballast, baffle, cavitation, chain rattle, trawl chain rattle, and seismic profile. Each anthropogenic transient subclass will be described below.

### 6.1.1 Ballast

Ballast sounds are associated with surface vessels – they are produced by wave motion pitching and rocking the ship. The water motion in the ballast tank produces ballast-

related transients. The example ballast sound shown in Figure 6.1 begins approximately at the two second point. There were a total of 24 ballast transients included in the dataset with a mean duration of 0.33 s and 0.05 s variance in duration ($\sigma^2$).

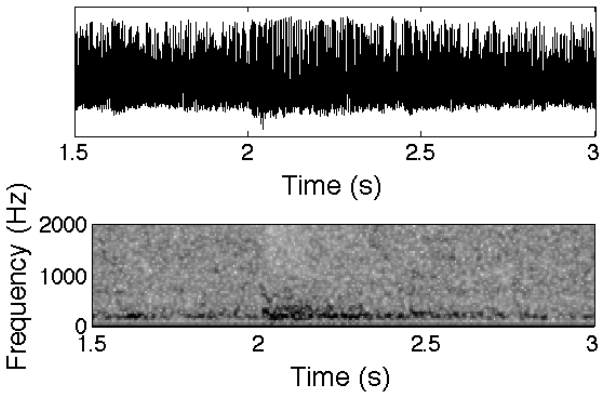


Figure 6.1 Time series and spectrogram of a ballast sound. The spectrogram was generated using a Hamming window length of 512 samples with an overlap of 80%.

### 6.1.2 Baffle

Like ballast sounds, baffle sounds are produced by wave motion pitching and rocking the ship. Baffle and ballast sounds are often found together. The baffle transients are generated when the water sloshes off the ballast tank baffles. There were 30 baffle sounds included in the dataset that have a mean duration of 0.14 s ($\sigma^2$ = 0.01 s). Multiple baffle sounds are shown in Figure 6.2.

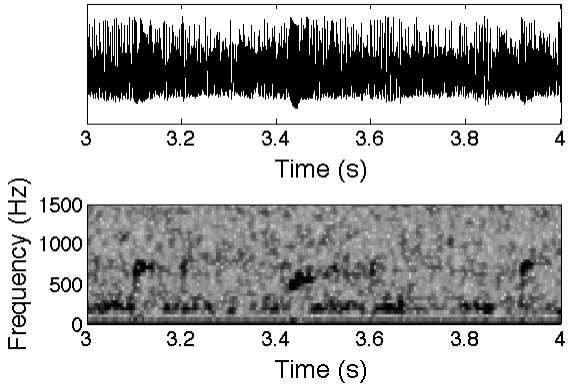


Figure 6.2 Time series and spectrogram of a baffle sound. The spectrogram was generated using a Hamming window length of 512 samples with an overlap of 80%.

### 6.1.3 Cavitation

Cavitation transients are associated with the movement of vessels and are generated by the collapse of bubbles produced by a vessel's propeller. An example of a cavitation event is shown in Figure 6.3. There were nine cavitation transients included in the dataset with mean duration 2.82 s ($\sigma^2$ = 0.49 s).

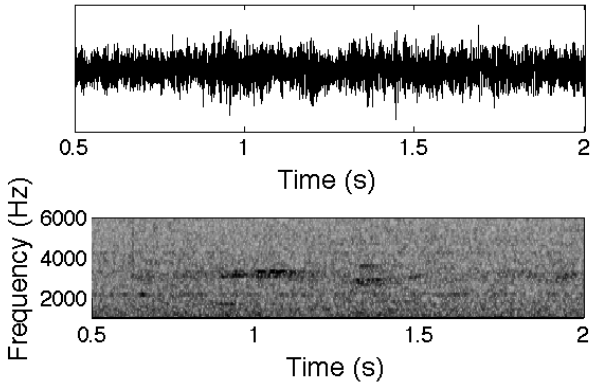


Figure 6.3 Time series and spectrogram of a sound generated by cavitation. The spectrogram was generated using a Hamming window length of 512 samples with an overlap of 70%.

### 6.1.4 Chain rattle

The chain rattle class was further subdivided into two groups – chain rattle and trawl chain rattle. Trawl chain rattles (Figure 6.4) result from fishing trawler activity and are

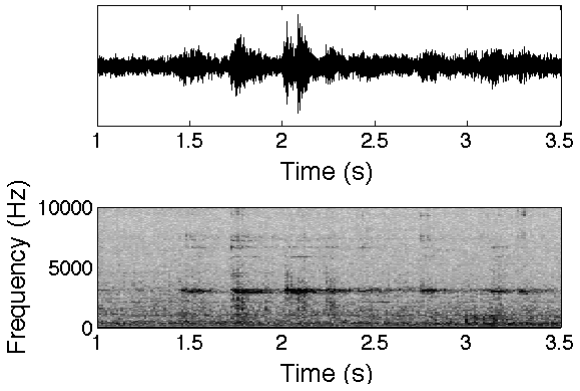


Figure 6.4 Time series and spectrogram of a sound generated by a trawl chain rattle. The spectrogram was generated using a Hamming window length of 1024 samples with an overlap of 50%.

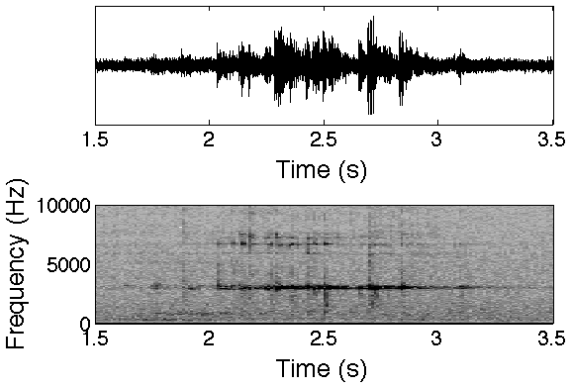


Figure 6.5 Time series and spectrogram of a sound generated by a chain rattle. The spectrogram was generated using a Hamming window length of 1024 samples with an overlap of 50%.

often used as an alert to submarine crews of a potential safety hazard. Any chain rattle sounds that could not be positively identified as resulting from a trawl chain were placed in the more general chain rattle category (Figure 6.5). Trawl chain rattle sounds had a mean duration of 0.36 s ($\sigma^2 = 0.13$) and chain rattle sounds had a mean duration of 0.52 s ($\sigma^2 = 0.28$). There were a total of 53 trawl chain rattle sounds and 52 general chain rattle sounds.

### 6.1.5 Seismic profile

Airguns used for subsurface analysis of the seabed produce seismic profile transients by

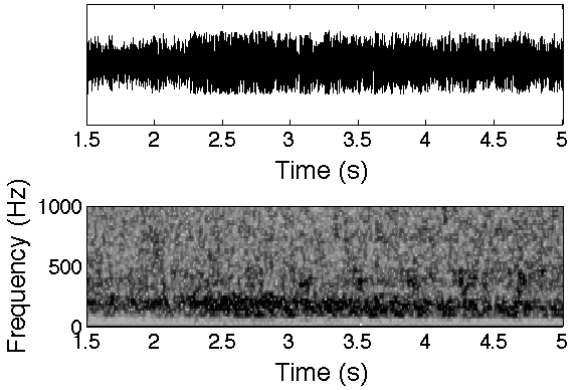


Figure 6.6 Time series and spectrogram of a seismic profile sound. The spectrogram was generated using a Hamming window length of 512 samples with an overlap of 80%.

releasing highly compressed air to produce sound energy; important geologic information can be obtained from the sound waves reflected from the seabed and strata layers [50]. There were a total of four seismic profile sounds included in the dataset. The mean duration of seismic profile sounds was 1.71 s ($\sigma^2 = 0.09$ s). An example seismic profile is shown in Figure 6.6.

### 6.1.6 Dataset Summary

A summary of the anthropogenic passive transient sounds (by sub-class) and sperm whale clicks, including number of each type of transient, is listed in Table 6.1. The sample rates of each class of transient varied based on the original recording source; to ensure consistent treatment of the sounds by the aural classifier each sound was resampled to 8 kHz using SoX [44]. All six types of anthropogenic transients were treated as a single class of sound to be classified against sperm whale clicks.

Table 6.1 Number of sounds by sub-class in the anthropogenic transient and sperm whale click dataset.

| **Sound Type** | **Number in Dataset** |
|---|---|
| Ballast | 24 |
| Baffle | 30 |
| Cavitation | 9 |
| Trawl chain rattle | 52 |
| Chain rattle | 53 |
| Seismic profile | 4 |
| Sperm whale clicks | 178 |

## 6.2 RESULTS AND DISCUSSION

As discussed in Section 5.3.1, it is difficult to determine the best number of features to select to include in the principal components based only on the cumulative Fisher score. Thus, for classification of sperm whale clicks and anthropogenic transients, classification was performed for all possible number of selected features. Figure 6.7 shows the *AUC* values for classification with 2 to 39 (all non-redundant) selected features. All numbers of selected features produced large *AUC* values; however, the maximum *AUC* value of

1.00 (rounding up) resulted from using twelve selected features. Classification with twelve features will be presented first since it is known that this classification model will produce excellent results. Note that all *AUC* values shown in Figure 6.7 are either 0.99 or 1.00 when represented with the appropriate number of significant figures; although all of the *AUC* values are not statistically different from each other, the general trend (i.e. the local minimum around 29 selected features) is, itself, interesting. Since differences in *AUC* value are only noticeable to three figures (there are only two significant figures), one might expect more random fluctuations in the curve; instead there is a local minimum that does not seem to be consistent with fluctuations due only to inaccuracy of measurement.

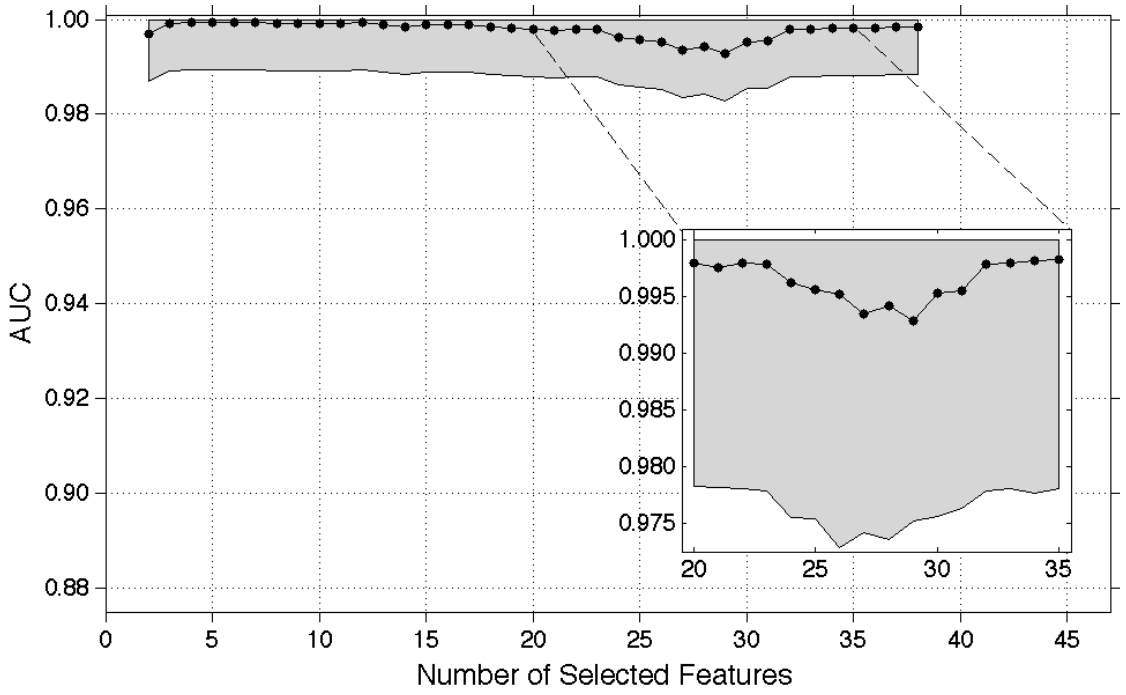


Figure 6.7 Performance results for classification of sperm whale clicks and anthropogenic transients with respect to number of features included in the principal components. A zoomed in view of the local minimum is displayed in the inset figure. The grey region represents the estimated error resulting from calculation of the *AUC*. Points are connected merely for visualization purposes and are not intended to imply that the data are continuous.

### 6.2.1 Classification with Twelve Selected Features

The decision region using twelve selected features is depicted in Figure 6.8. The aural classifier was very effective at discriminating between sperm whale clicks and the anthropogenic transients – the classification accuracy was 98% with all sperm whale clicks correctly classified and only anthropogenic transients misclassified. The corresponding ROC curve (the blue curve) is displayed in Figure 6.9. The *AUC* is rounded up to 1.00 and the equal error rate is 1%, which taken together all indicate near-ideal classification performance. The value $p_2 = 0.73$ demonstrates that a large amount of the variance in the twelve selected features is represented in the first two principal components. As was seen in the multiclass decision regions in section 5.1 that included sperm whale clicks, the sperm whale clicks are spread along a line in the PCA space. The covariance and correlation of the sperm whale click data points are -1.75 and -0.92, respectively. Interestingly, the anthropogenic transients also follow a linear trend with a covariance of 2.62 and correlation of 0.97.

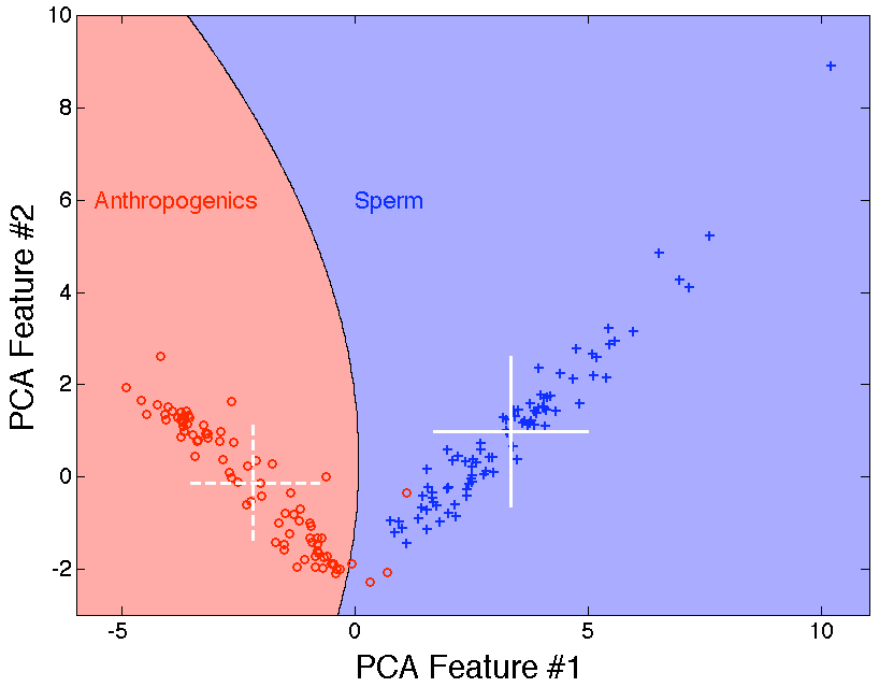


Figure 6.8 Decision region for binary classification of sperm whale clicks and anthropogenic transients. Classification was performed with twelve selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

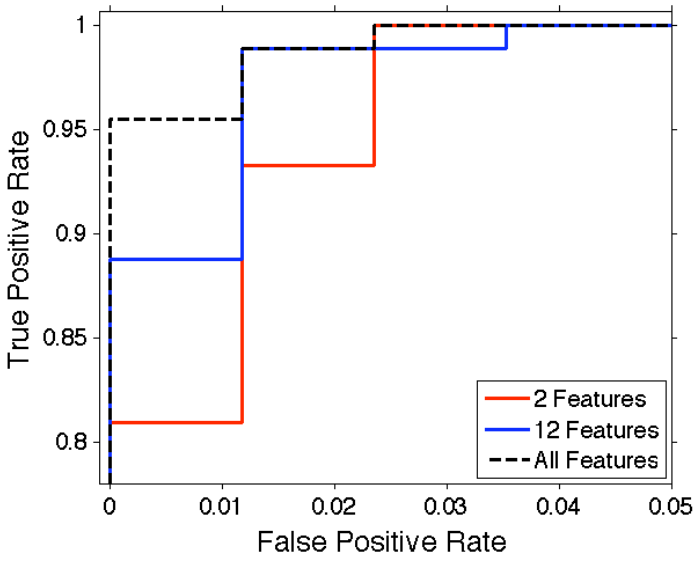


Figure 6.9 Sperm whale and anthropogenic transient ROC curves for classification with 2,12, and 39 features. Only the region where the ROC curves do not overlap is plotted. These ROC curves correspond to the decision regions shown in Figure 6.8, Figure 6.10, and Figure 6.14.

### 6.2.2 Classification with Two Selected Features

The linear arrangement of data points in both classes was investigated further by performing classification with only the two features with the largest Fisher scores – loudness centroid and global mean sub-band decay slope. Two features are used because it should be relatively simple to establish a relationship between the selected features and the principal components. The resulting decision region is shown in Figure 6.10. The same linear within-class trend observed in Figure 6.8 is also evident in this PCA space. The correlation and covariance of the sperm whale click data points are -0.43 and -0.95, respectively. The anthropogenic transient data points have a covariance of 0.89 and correlation of 0.99. With only two features selected the aural classifier still performed well, with an accuracy of 98%, $AUC = 1.00$ (rounded up, see red curve in Figure 6.9), and equal error rate of 2%. All three of the performance measures are indicative of near-ideal classification results.

The normalized feature values of the same two features that were used in PCA to form the decision region in Figure 6.10 are plotted in Figure 6.11. This plot represents the

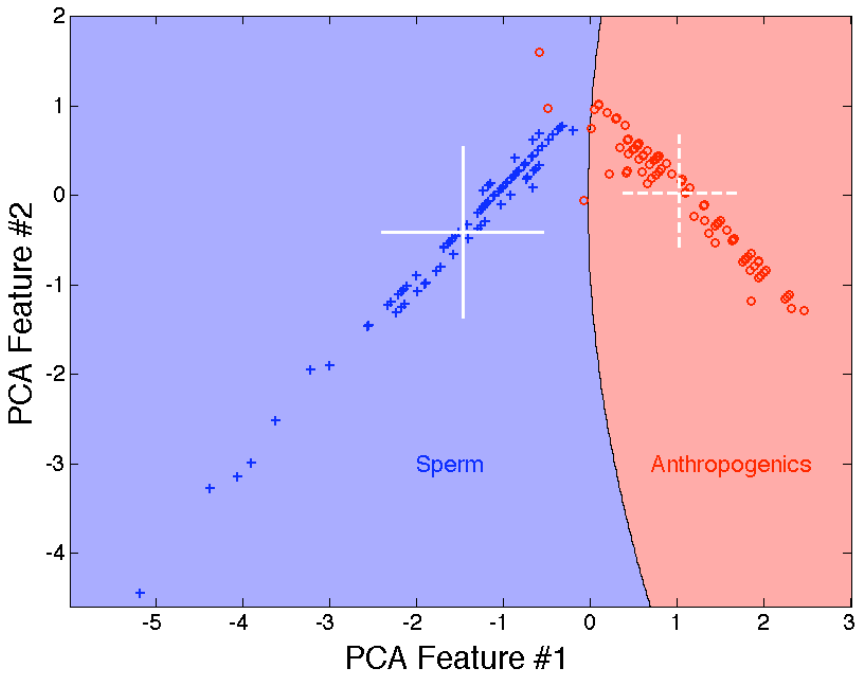


Figure 6.10 Decision region for binary classification of sperm whale clicks and anthropogenic transients. Classification was performed with two selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

relationship between feature values prior to performing PCA. Histograms of the loudness centroid and global mean sub-band decay slope are shown in Figure 6.12 and Figure 6.13. Together these plots show that all sperm whale clicks have similar loudness centroids but a wide range of values for the global mean sub-band decay slope. Conversely, anthropogenic transients have a wide range of loudness centroids but relatively similar values for global mean sub-band decay slope. The within-class variances of these two feature values are listed in Table 6.2. For each feature, the within-class variance for one class is small relative to the other class.

Table 6.2 With-in class variance of the normalized loudness centroid and global mean sub-band decay slope features (before PCA).

| | **Sperm Whale** | **Anthropogenics** |
|---|---|---|
| **Loudness Centroid** | 0.009 | 0.95 |
| **Global mean sub-band decay slope** | 0.93 | 0.19 |

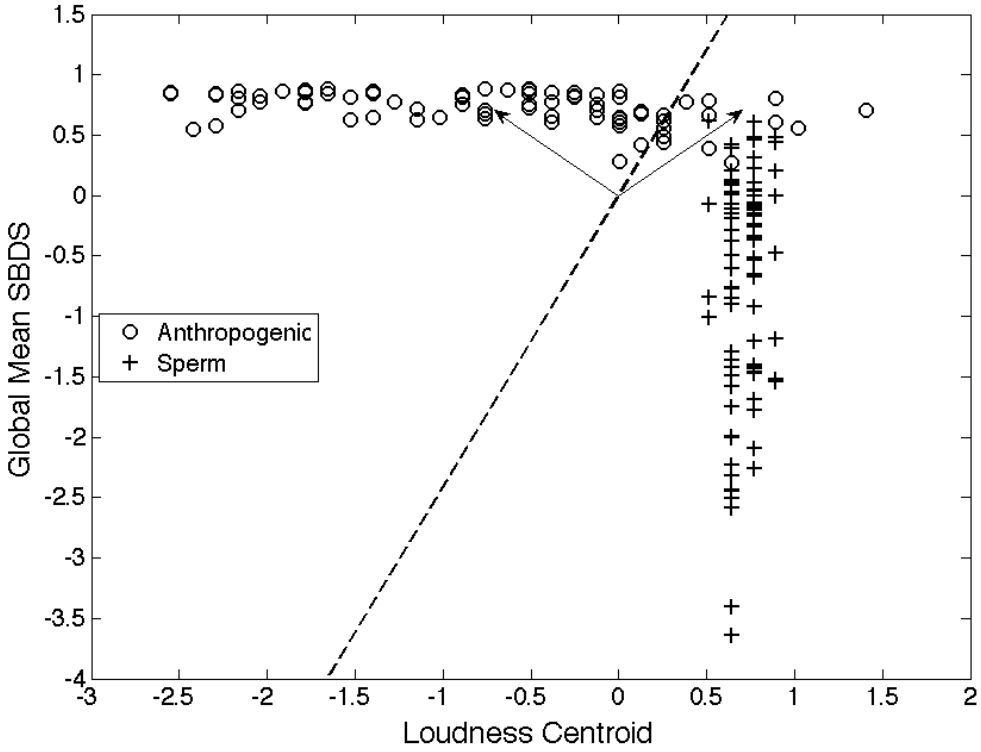


Figure 6.11 The two features ranked highest by the Fisher score. These results were plotted before performing PCA on the selected features. The two arrows represent the principal components and the dotted line is the line through which data points are reflected when PCA is performed.

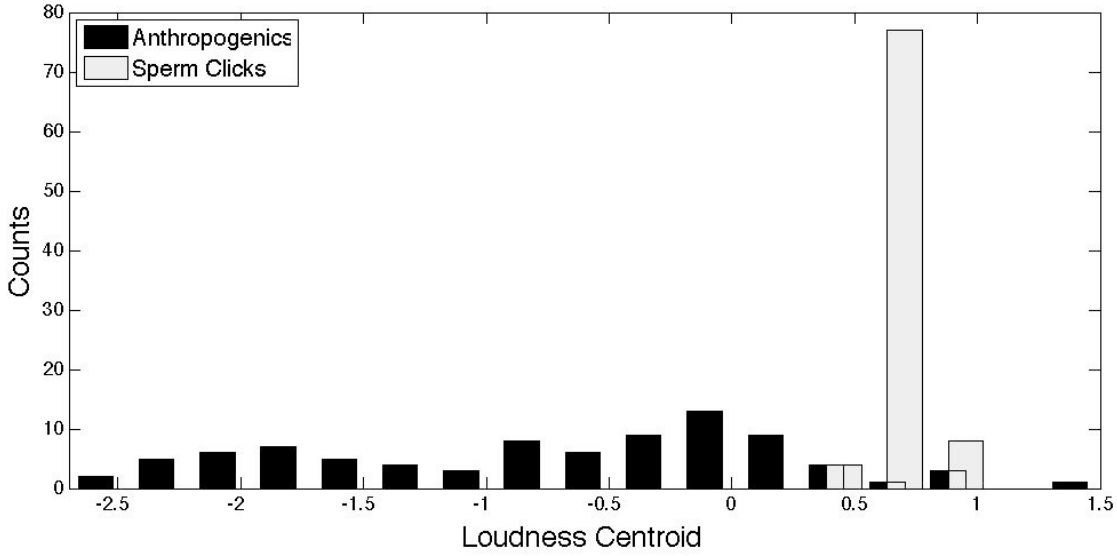


Figure 6.12 Histogram of loudness centroid values, the highest ranked feature for discriminating between sperm whale clicks and anthropogenic transients.

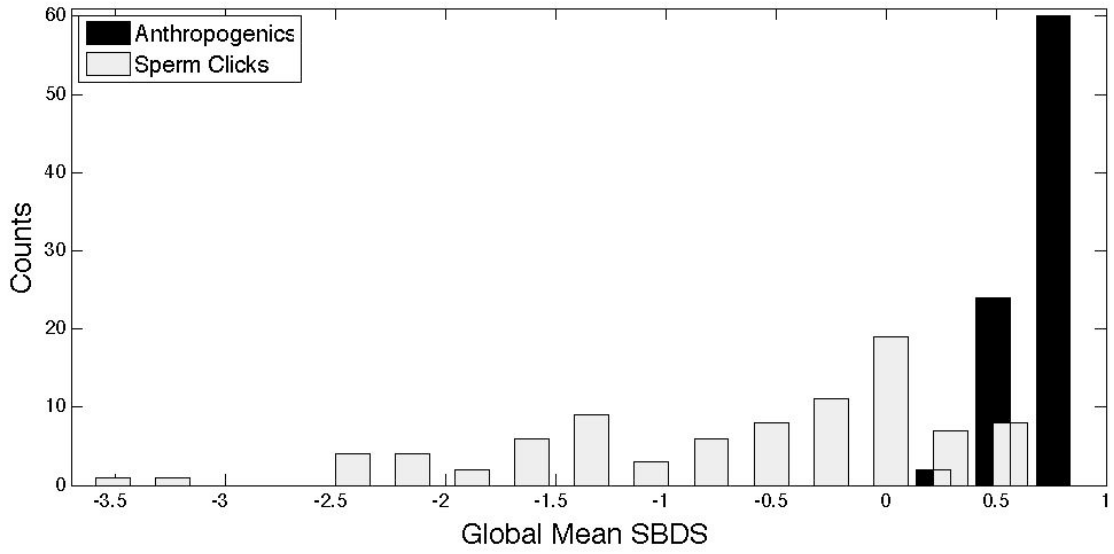


Figure 6.13 Histogram of global mean sub-band decay slope (SBDS) values, the second highest ranked feature for discriminating between sperm whale clicks and anthropogenic transients.

As a side note, the case where two features are selected is a relatively simple example that can be used to demonstrate the results of PCA (refer to Section 2.1.5 for the theory of PCA). The eigenvectors of the covariance matrix, also referred to as the principal components, form the transformation matrix. In this case the transformation matrix is composed of the two principal components, $\mathbf{a}_1 = \frac{1}{2}\begin{bmatrix}-\sqrt{2}\\ \sqrt{2}\end{bmatrix}$ and $\mathbf{a}_2 = \frac{1}{2}\begin{bmatrix}\sqrt{2}\\ \sqrt{2}\end{bmatrix}$. Although the eigenvector elements are shown to be $\sqrt{2}/2$, the computer-based aural classification program performs computations with only finite precision; however, the estimate that the elements of the eigenvectors are $\sqrt{2}/2$ is accurate to at least eight decimal places. Therefore, this example will continue to use the compact $\sqrt{2}/2$ notation. The first principal component, $\mathbf{a}_1$, lies in the direction of maximum variance and the second principal component, $\mathbf{a}_2$ is orthogonal to $\mathbf{a}_1$. The transformation matrix $\mathbf{A}$, composed of $\mathbf{a}_1$ and $\mathbf{a}_2$ is thus written as,

$$\mathbf{A} = \frac{1}{2}\begin{bmatrix}-\sqrt{2} & \sqrt{2}\\ \sqrt{2} & \sqrt{2}\end{bmatrix} \quad . \qquad \text{Eqn. 6.1}$$

The matrix **A** has a determinant of -1, indicating that the linear transformation results in a reflection through a line [51] – the line of reflection be found by solving the following system of equations:

$$\begin{aligned} x &= \tfrac{1}{2}\left(-\sqrt{2}x + \sqrt{2}y\right) \\ y &= \tfrac{1}{2}\left(\sqrt{2}x + \sqrt{2}y\right) \end{aligned} \quad . \qquad \text{Eqn. 6.2}$$

The solution to these equations is $y = \left(\sqrt{2}+1\right)x$, which corresponds to the reflection line plotted in Figure 6.11. This example provides a clue as to how the linear arrangement of data points in the PCA space comes about.

### 6.2.3 Classification with All Non-redundant Features

The decision region shown in Figure 6.14 was generated using all 39 non-redundant features. Classification was completed with 98% accuracy. The ROC curve corresponding to this classifier model is shown in Figure 6.9 with $AUC = 1.00$ (rounded up) and an equal error rate of 1%. In this case, classification performance did not

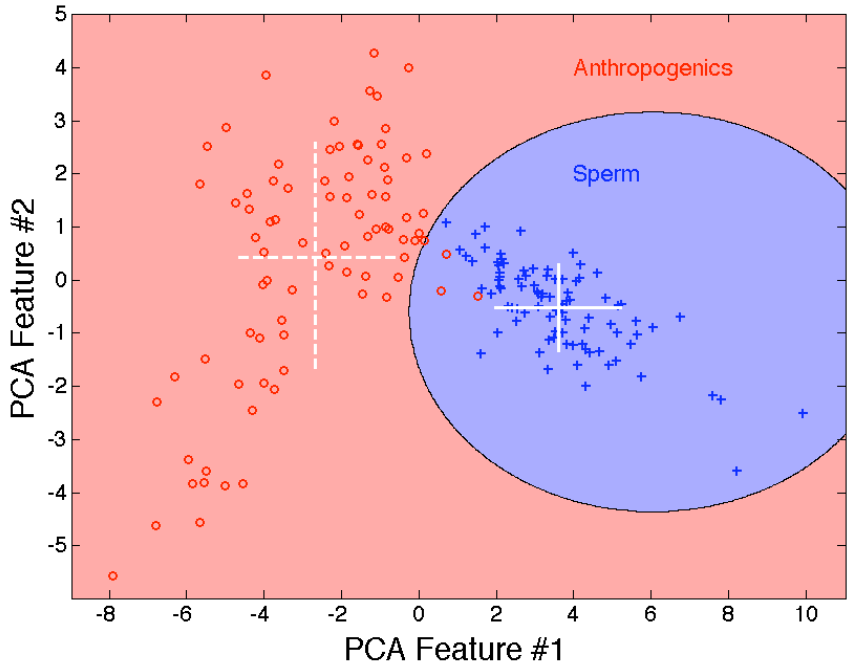


Figure 6.14 Decision region for binary classification of sperm whale clicks and anthropogenic transients. Classification was performed with all 39 non-redundant features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

noticeably change with number of selected features; however, it is apparent from the decision region that when more features were included in the principal components the linear trend previously observed in both the anthropogenic transient and sperm whale click classes is no longer present. When all non-redundant features were used the within-class covariance for the sperm whale clicks was 2.64 and the correlation was 0.58. The covariance and correlations of the anthropogenic transients were -0.98 and -0.73, respectively. Though the covariance and correlation for each class are non-zero, incorporating additional features significantly reduced these values, and thus the linear within-class trend is no longer as apparent. The other features that were included provided additional information that allowed the principal component analysis transformation to distribute the data points more evenly in all directions around the class mean.

### 6.2.4 Comparison of Classification Results

The three features ranked highest by PCA are listed in Table 6.3 and weighting of all features in the principal components using 12 and 39 features is represented in Figure 6.15. Most of the 12 selected features received similar weighting to each other, whereas some of the 39 non-redundant features received much larger weight values than other features in the corresponding principal components. Many of the features that were highly ranked when only 12 features were selected, were ranked low in the principal components composed of all non-redundant features. All but one of the features with high weighting in the principal components are time-frequency features. Most of these

Table 6.3 Three highest weighted features using 2 and 12 selected features, and all non-redundant for sperm whale and anthropogenic transient classification.

| **2 Selected Features** | **12 Selected Features** | **All Features** |
|---|---|---|
| Loudness centroid | Local minimum sub-band decay slope | Global maximum sub-band decay time |
| Global mean sub-band decay slope | Local maximum sub-band attack slope | Global minimum sub-band attack time |
| | Global mean sub-band attack slope | Frequency of global minimum sub-band attack slope |

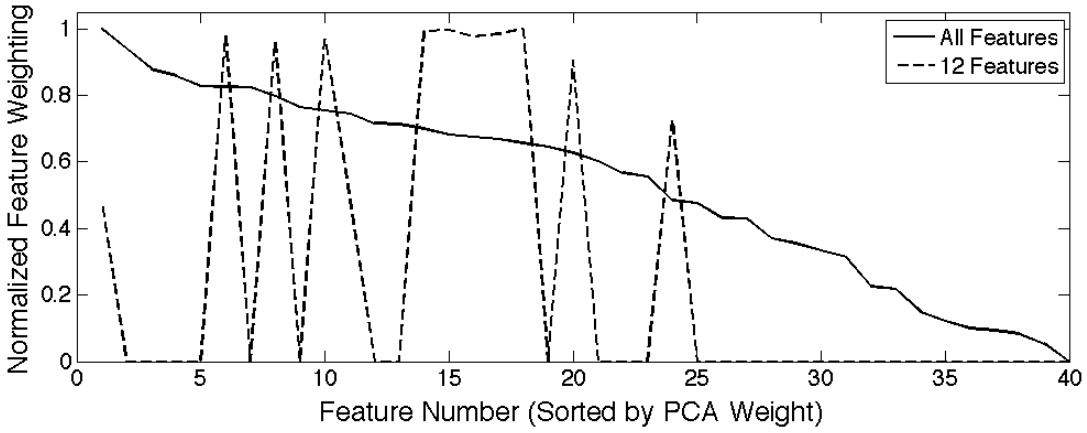


Figure 6.15 Normalized weighting of features in the first two principal components. Features are sorted from largest PCA feature weighting to smallest based on PCA with all 39 non-redundant features. These eigenvectors correspond to the decision regions shown in Figure 6.8 and Figure 6.14. Peaks are connected merely for visualization purposes and are not intended to imply that the data are continuous.

features describe the rise (i.e. attack) or decay slope of the sounds. In general, the rise times and decay times of the sperm whale clicks are less than for anthropogenic transients. Anthropogenic transients had larger decay time values because there is generally more reverberation audible after the anthropogenic transients than for sperm whale clicks; additionally, anthropogenic transients typically were of longer duration.

The decision regions using 2, 12, and 39 features are depicted in Figure 6.16 with the anthropogenic transient subclasses each represented by their own symbol. Samples from the trawl chain rattle and chain rattle classes were the only anthropogenic transients to be misclassified as sperm whale clicks. The chain rattles have similar characteristics to sperm whale clicks: they are more broadband and have shorter duration than many of the other anthropogenic transients, and at some frequencies have large rise and decay slopes. There appears to be a general clustering of trawl chain rattles with chain rattles in the decision regions. Other anthropogenic transient subclasses are more disperse in the decision regions, possibly because there are fewer samples in the dataset with which to accurately represent the distribution of the subclasses.

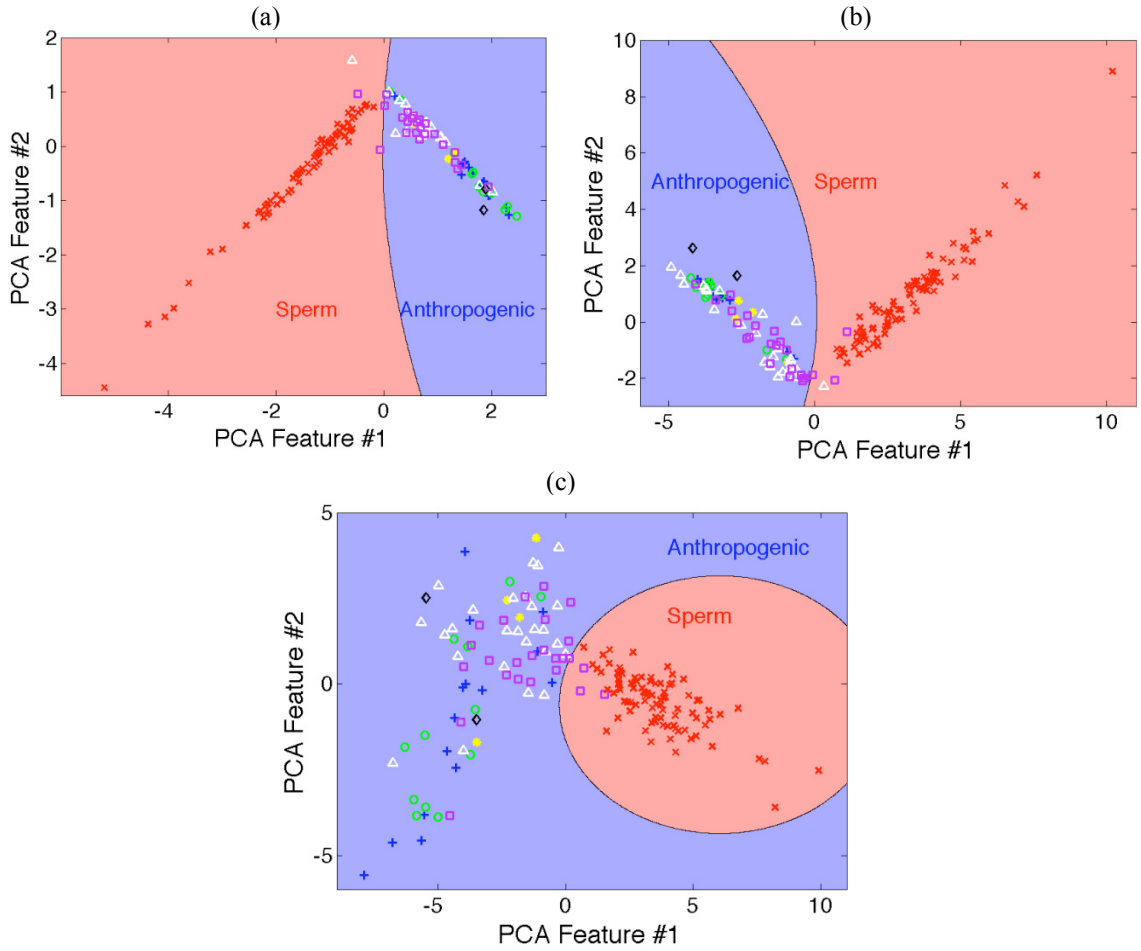


Figure 6.16 Decision regions for binary classification of sperm whale clicks and anthropogenic transients. These decision regions are identical to those shown in Figure 6.8, Figure 6.10, and Figure 6.14, except in this figure the subclasses of anthropogenic transients are each represented by their own symbol: baffle (blue cross), ballast (green circle), cavitation (yellow star), chain rattle (white triangle), seismic profile (black diamond), and trawl chain rattle (purple square). Classification was performed with (a) two features, (b) twelve features, and (c) all 39 non-redundant features.

### 6.2.5 Comparison of Sperm Whale Classification with Literature Results

Kandia and Stylianou [52] used the Teager-Kaiser energy operator to automatically detect sperm whale clicks in acoustics recordings. They compared Teager-Kaiser detection results to detection results from the commercially available Rainbow Click detection suite. The true positive detection rate for the Teager-Kaiser energy operator method was 94.05% and for the Rainbow Click detector was 71.12%. When presented with previously detected signals, the aural classifier was able to correctly identify 98% of

the signals, which is significantly better than the Rainbow Click detector results and about four percentage points better than the Teager-Kaiser energy operator method. Kandia and Stylianou proposed that an advantage of their method is that their algorithm requires few input parameters to perform detection; this is also an advantage of the aural classifier because it requires few user-defined inputs.

Huynh *et al.* [53] compared classification results of four species of cetaceans – sperm whales, dolphins, and porpoises – and noise (a total of four classes) using two different feature extraction methods. When features were extracted from the Fourier transform of the signals, sperm whale clicks were classified with only 57.5% accuracy. When the wavelet transform was used for feature extraction, classification accuracy of sperm whale clicks increased to 98.5%. Thus, a classifier that used features extracted from the wavelet transform produced similar results to classification with aural features although the perceptual features performed significantly better than features extracted directly from the Fourier transform.

## 6.3 CONCLUSIONS

The aural classifier was able to accurately discriminate between sperm whale clicks and a variety of anthropogenic transients, even though the subclasses of anthropogenic transients had distinct aural characteristics. The maximum classification performance was obtained when twelve features were used for PCA – classification accuracy was 98%, the *AUC* was 1.00 and the equal error rate was 1%. Excellent classification performance results were observed for all numbers of selected features, ranging between 2 and 39 (all non-redundant features) features. The lowest noted *AUC* value was 0.99, which is still indicative of a successful classification.

Time-frequency features were especially important for classification of sperm whale clicks and anthropogenic transients. In general, the anthropogenic transients produced smaller values of the attack and decay slopes than sperm whale clicks. The notably different time-frequency features between classes may be explained by the lack of

audible reverberation and the quick energy transition (i.e. rapid rise time), from ambient noise to signal, in sperm whale clicks.

A linear pattern of data points within the sperm whale and anthropogenic passive transient classes was observed when small numbers of features were selected. This trend was analyzed by performing classification with only two features and examining the within-class variances of features that were highly ranked in the principal components. The trend was attributed to relatively small within-class variance for highly ranked features for one of the classes compared to relatively large within-class variance for the other class. The results of this analysis can be applied to the linear trend within the sperm whale class in the multiclass decision regions presented in Section 5.1.1.

# CHAPTER 7 DISCRIMINANT ANALYSIS IMPLEMENTATION

High dimensionality feature spaces, such as the 58-dimensional space used by the aural classifier, may not be adequately represented when using a limited dataset. Using too many dimensions to describe a feature space is analogous to overfitting a polynomial. For example, given ten data points generated by adding noise to a quadratic equation, it is possible to perfectly fit a $10^{th}$ degree polynomial to these points; however, it does not accurately represent the data because predictions for new data points are not likely to be similar to data points originating from the underlying quadratic nature of the dataset [24]. In the case of the aural classifier, the features were selected because of the belief that each one may improve discrimination between (at least) some of the classes. If it is assumed that all of the features improve discrimination, then a method must be used to find the best combination of features for discrimination while reducing the dimensionality of the feature space.

As discussed in Section 2.1.5, the current implementation of the aural classifier uses principal component analysis (PCA) to project the multi-dimensional space defined by the selected features (see Section 2.1.4 for description of how features are selected) onto a lower dimensional space in which the resulting axes are linear combinations of the selected features [9]. Two principal components are typically selected, mainly due to the ease with which two-dimensional data can be represented graphically; however, one must be aware that for classification to be successful in the lower dimensional space a relatively large amount of the variation of the higher-dimensional dataset should be

maintained. The advantage of PCA is that it produces a projection that best *represents* the data in a least-squares sense [24], i.e., PCA does the best job of maintaining the most scatter/variation of the dataset.

Although PCA finds the feature combinations that are useful for representing the data, there is no reason to assume that these are the best components to discriminate between classes. When attempting to discriminate between classes, it is likely better to seek a projection that best *separates* the classes in a least squares sense. The method that accomplishes this is discriminant analysis (DA). The aim of DA is to find the combination of features that results in a projection containing the most distance between class means relative to the standard deviations of the classes [24]. The theory of discriminant analysis was presented in Section 2.1.6. It is hypothesized that implementing DA should improve the ability of the classifier to discriminate between classes. The remainder of this chapter investigates if replacing PCA with DA will improve classification results.

## 7.1 Comparison and Discussion of DA and PCA Results

Comparison of classification results using DA versus PCA was accomplished using the cetacean dataset. Possible effects of number of classes ($c$) on classification were also considered by performing classification with two, three, and five cetacean classes. These test cases were selected because there are a total of *five* classes in the dataset; when using DA with *three* classes, a maximum of two discriminant functions are produced that naturally define a 2D decision region without the need to eliminate any computed discriminant functions; and *two* class classification is the simplest case. The same training/testing split was used in all cases. PCA was performed in each case with the twenty best features, as selected by the Fisher Linear Discriminant score. Classification results using PCA and DA for feature space reduction are presented and discussed in the following sections.

### 7.1.1 Five classes ($c$ = 5)

All five cetacean species – bowhead, humpback, North Atlantic right, minke, and sperm whales – were included for comparison of automatic classification results when using either PCA or DA for feature space reduction. The results from applying PCA are discussed first.

The decision region generated using twenty features and two principal components is shown in Figure 7.1. Classification of vocalizations was 78% accurate. Table 7.1 contains the confusion matrix of *AUC* values corresponding to the decision region below – note that $M = 0.97$ for this case, indicative of a successful classification. Most of the overlap of data points is observed among the baleen whale species, especially the bowhead/humpback and right/minke pairs (see corresponding *AUC* values in Table 7.1). No sperm whale clicks were misclassified, although the data points spread out along a

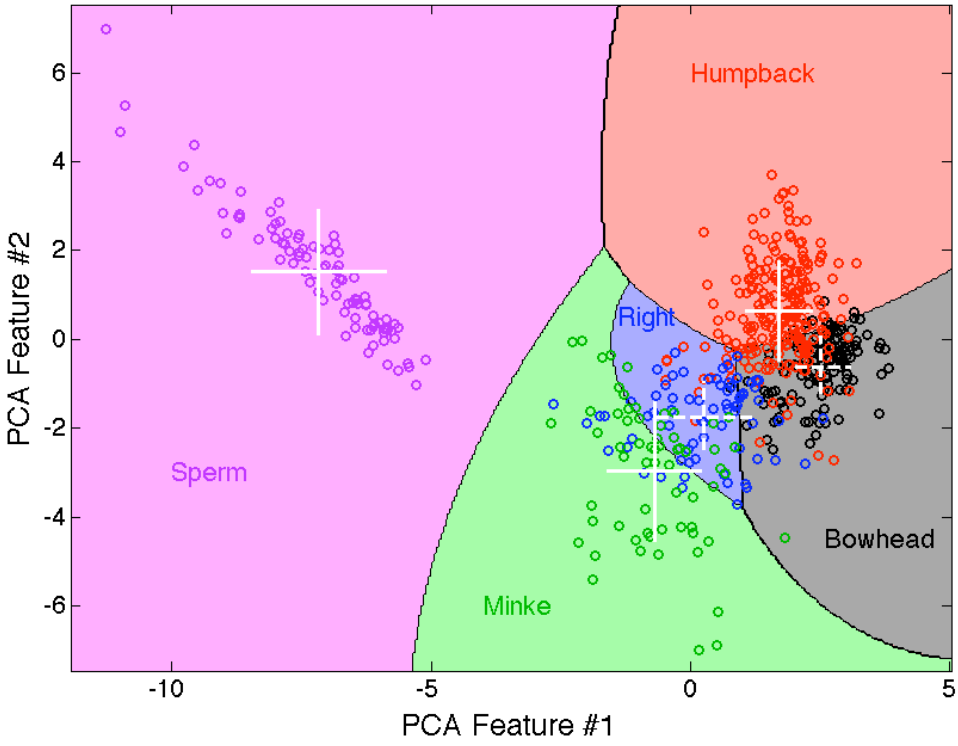


Figure 7.1 Decision region for classification of bowhead, humpback, right, minke, and sperm whale vocalizations. Data points from the testing subset were projected onto the 2D space using PCA on twenty selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

Table 7.1 Confusion matrix of *AUC* values corresponding to the decision region shown in Figure 7.1, where feature space dimensionality reduction is performed using PCA. The value $M = 0.97$. The asterisk indicates *AUC* values that appear to result from an ideal classifier, but in fact resulted from rounding to two decimal places.

| | **Humpback** | **Right** | **Minke** | **Sperm** |
|---|---|---|---|---|
| **Bowhead** | 0.89 | 0.98 | 1.00 | 1.00 |
| **Humpback** | | 0.98 | 1.00* | 1.00 |
| **Right** | | | 0.87 | 1.00 |
| **Minke** | | | | 1.00 |

line in the PCA space rather than forming a diffuse cluster of data points like the other classes (refer to Section 5.3.2 for a possible explanation of this trend). The PCA feature values for sperm whale clicks are strongly correlated between the two principal components, whereas there appears to be negligible correlation of the within-class scatter of the other cetacean species. The relative amount of variability contained in the first two principal components was calculated using Eqn. 2.10: $p_2 = 0.67$, that is, 67% of the scatter in the dataset is maintained by using two principal components.

Classification using DA for feature space reduction was carried out twice. The first classification was performed so that all four discriminant functions were used; thus, classification was done in the resulting four-dimensional space. Using all four discriminant functions is analogous to performing classification using all 20 principal components. The second classification was done using only the first two discriminant functions to facilitate visualization of the resulting decision regions – as is often done for PCA classification. Eqn. 2.21 can be used to determine the relative effectiveness of using a subset of discriminant functions – in the case shown in Figure 7.2 where the best two discriminant functions were used, $p_2 = 0.77$. Thus, a large percentage of the discriminability of the DA method is maintained when employing a subset of two discriminant functions.

Results obtained from using four discriminant functions for classification of vocalizations belonging to the testing subset are presented as the confusion matrix of pairwise *AUC* values (in Table 7.2). The accuracy of this classification was 86%. Classification results

are either perfect or near-perfect for all pairs of cetacean species except for bowhead/ humpback whales, for which $AUC = 0.91$. The pairwise $AUC$ was 1.00 for the bowhead/humpback pair in the training subset, which is noticeably greater than for the testing subset; since there is significant variability – particularly in the humpback vocalizations – a decision boundary may have been selected that was too specific to the training dataset to work well with the testing subset. Although a lower $AUC$ value is measured for the testing subset of bowhead/humpback vocalizations, it still represents a 91% chance of correctly discriminating between bowhead and humpback vocalizations, which is considered to be successful for these aurally similar vocalizations. Applying the aural classifier with all four discriminant functions was effective as can be noted by the large value of $M = 0.99$.

Table 7.2 Confusion matrix of $AUC$ values from the testing subset, corresponding to using the DA feature space dimensionality reduction method. Four discriminant functions were produced. The value $M = 0.99$. The asterisk indicates $AUC$ values that appear to result from an ideal classifier, but in fact resulted from rounding to two decimal places.

| | **Humpback** | **Right** | **Minke** | **Sperm** |
|---|---|---|---|---|
| **Bowhead** | 0.91 | 1.00* | 1.00 | 1.00 |
| **Humpback** | | 1.00* | 1.00 | 1.00 |
| **Right** | | | 1.00* | 1.00 |
| **Minke** | | | | 1.00 |

The decision region, produced by the two discriminant functions corresponding to the largest two eigenvalues, is represented in Figure 7.2 and the corresponding confusion matrix of $AUC$ values is given in Table 7.3. In this case $M = 0.98$ and the total accuracy of classification was 83%. As with the four-dimensional case, the most overlap between classes occurred for the bowhead/humpback pair. All sperm whale clicks were correctly classified. Limiting the number of discriminant functions to two reduced average classification results by only 0.005, as measured by the $M$-value – this is a statistically insignificant amount. Taking into account the ease of visualization in 2D, using only two discriminant functions in this case is preferred since a negligible amount of discriminability was lost. The accuracy decreased by only three percentage points when

only two discriminant functions were used, but remained more accurate than classification using PCA by five percentage points.

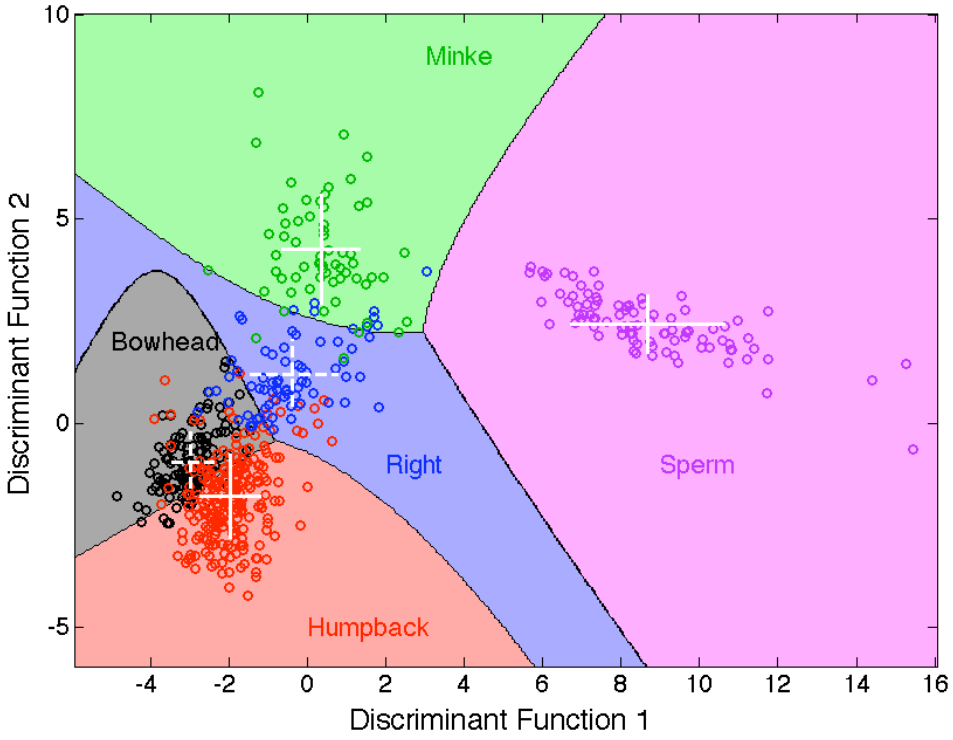


Figure 7.2 Decision region for classification of bowhead, humpback, right, minke, and sperm whale vocalizations. Data points from the testing subset were projected onto the 2D space using DA. When three classes are used, DA produces four discriminant functions; to allow plotting in 2D, the discriminant functions corresponding to the two largest eigenvalues were used for classification. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

Table 7.3 Confusion matrix of *AUC* values corresponding to the decision region shown in Figure 7.2, where feature space dimensionality reduction is performed using DA. The value $M = 0.98$.

| | **Humpback** | **Right** | **Minke** | **Sperm** |
|---|---|---|---|---|
| **Bowhead** | 0.90 | 0.99 | 1.00 | 1.00 |
| **Humpback** | | 0.98 | 1.00 | 1.00 |
| **Right** | | | 0.97 | 1.00 |
| **Minke** | | | | 1.00 |

Classification results in the four-dimensional and two-dimensional DA reduced feature spaces are slightly better than in the PCA space ($\Delta M = 0.02$ and $\Delta M = 0.01$, respectively). By visually comparing the two decision regions shown in Figure 7.1 and Figure 7.2 the

theoretical differences between PCA and DA methods are evident. The within-class scatter was noticeably reduced when discriminant functions were used; for example, the sperm whale data points form a tighter cluster around the class mean in the DA space but are more spread out in the PCA space. Increased distances between class means – when using DA rather than PCA – is also another evident difference between dimension reduction methods. Table 7.4 contains summary statistics on the separation between class means. Both of the dimensionality reduction methods caused all sperm whale clicks to be correctly classified, although the class mean of sperm clicks was further removed from the other class means when using DA.

Table 7.4 Summary statistics describing the distance between class means when reduced feature spaces are composed of either principal components or discriminant functions. The five class means correspond to the white crosses displayed on Figure 7.1 and Figure 7.2.

| | **2 Principal Components (All Species)** | **2 Discriminant Functions (All Species)** | **4 Discriminant Functions (All Species)** |
|---|---|---|---|
| **Minimum distance** | 0.57 | 1.03 | 2.57 |
| **Maximum distance** | 8.34 | 10.25 | 15.01 |
| **Mean distance** | 3.85 | 4.56 | 8.16 |

Examining the eigenvectors used for PCA and DA (with four discriminant functions) exposes the weighting of each feature in the resulting decision space. For example, the normalized sum of eigenvector components corresponding to the integrated loudness feature for the principal components is 1.00 compared to 0.70 for the discriminant functions; it can thus be concluded that integrated loudness is more important for the PCA method than for the DA method. Comparing eigenvectors allows one to determine the most important features for either the PCA or DA methods and which features are useful for both methods. By examining the feature weightings represented in Figure 7.3 it is evident that the twenty features selected for PCA correspond to features that have relatively large weights in the DA method – this is because the Fisher discriminant score is used to select the features that will form the principal components. Therefore, the selection of features for use in PCA is done in an analogous way to how features are weighted in the discriminant functions; however, the PCA method does weight features

differently than the DA method – as can be seen from examining the relative weighting values for particular features. Many of the PCA features have similar weighting to other features in the principal components, but different weightings than in the discriminant functions. The three highest weighted features for each of the dimension reduction methods is listed in Table 7.5. It is clear that many of the most important features for capturing the variance in the dataset are not the same as the features that best separate the classes.

Table 7.5 Three highest weighted features using PCA and DA methods.

| **PCA Features** | **DA Features** |
|---|---|
| Integrated loudness | Peak loudness value |
| Psychoacoustic bin-to-bin difference | Psychoacoustic bin-to-bin difference |
| Pre-attack peak loudness value | Mean sub-band correlation |

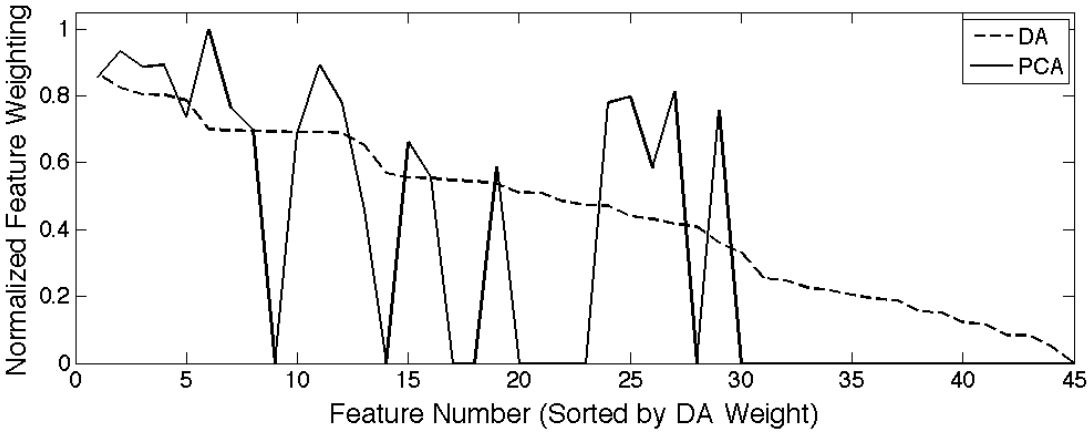


Figure 7.3 Normalized weighting of features for classification when using either PCA or DA for dimensionality reduction during classification of all cetacean species. Features are sorted from largest DA feature weighting to smallest. These eigenvectors correspond to the PCA-based decision region shown in Figure 7.1 and DA with four discriminant functions (results listed in Table 7.2). Peaks are connected merely for visualization purposes and are not intended to imply that the data are continuous.

### 7.1.2 Three classes ($c$ = 3)

For this analysis the three species selected were bowhead, humpback and right whales. Data points for these species exhibited the most overlap during classification of all

species (see Section 7.1.1) consistent with the fact that they are the most aurally similar. Additionally, in the case of three classes, a maximum of two discriminant functions can be calculated, thus removing the need to choose a subset of discriminant functions when plotting data in two dimensions.

The decision region formed using twenty selected features and two principal components is shown in Figure 7.4 and the corresponding confusion matrix of pairwise *AUC* values appears in Table 7.6. Average classification performance is good with $M = 0.98$ and an accuracy of 86%. There was more overlap with the bowhead/humpback pair than with other pairs of species and there was no difficulty discriminating between the North Atlantic right whale and humpback vocalizations. Each of the species exhibited relatively broad within-class scatter.

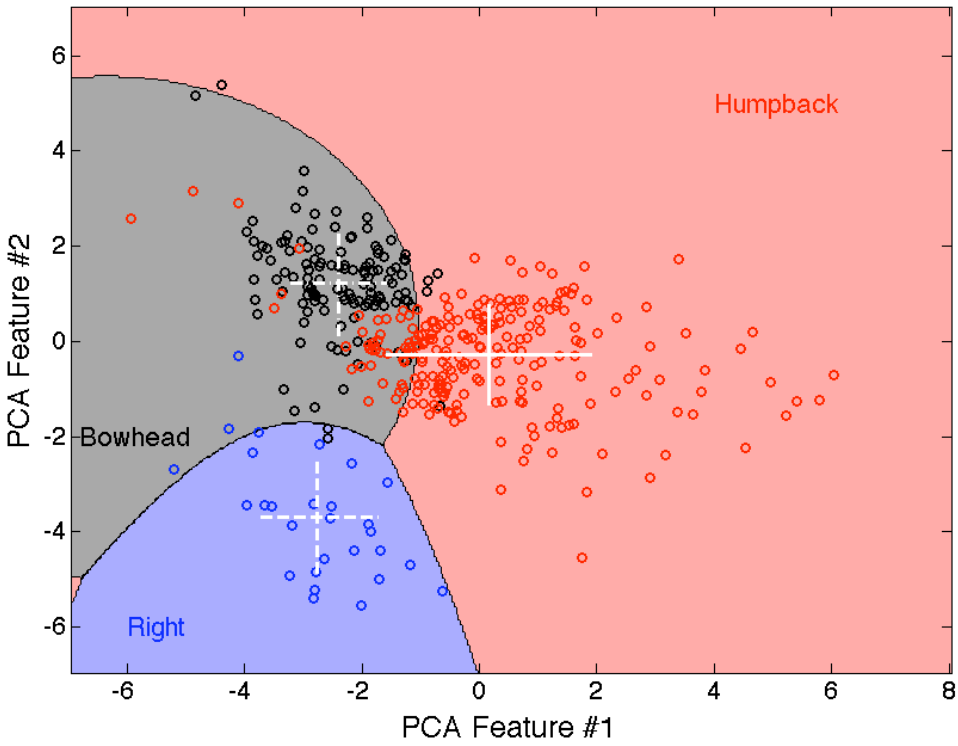


Figure 7.4 Decision region for classification of bowhead, humpback and right whale vocalizations. Data points from the testing subset were projected onto the 2D space using PCA on the selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

Table 7.6 Confusion matrix of *AUC* values corresponding to the decision region shown in Figure 7.4, where feature space dimensionality reduction is performed using PCA. $M = 0.98$.

| | **Humpback** | **Right** |
|---|---|---|
| **Bowhead** | 0.93 | 0.99 |
| **Humpback** | | 1.00 |

The two discriminant functions produced using DA on the three baleen whale species were used to project data onto the decision region shown in Figure 7.5. Table 7.7 contains the confusion matrix of *AUC* values. An *M*-value of 0.98 was calculated from the classification results, indicating successful classification. Classification was completed with an accuracy of 84%. Again, most of the misclassifications resulted from the bowhead/humpback pair, whereas very few right whale vocalizations were misclassified.

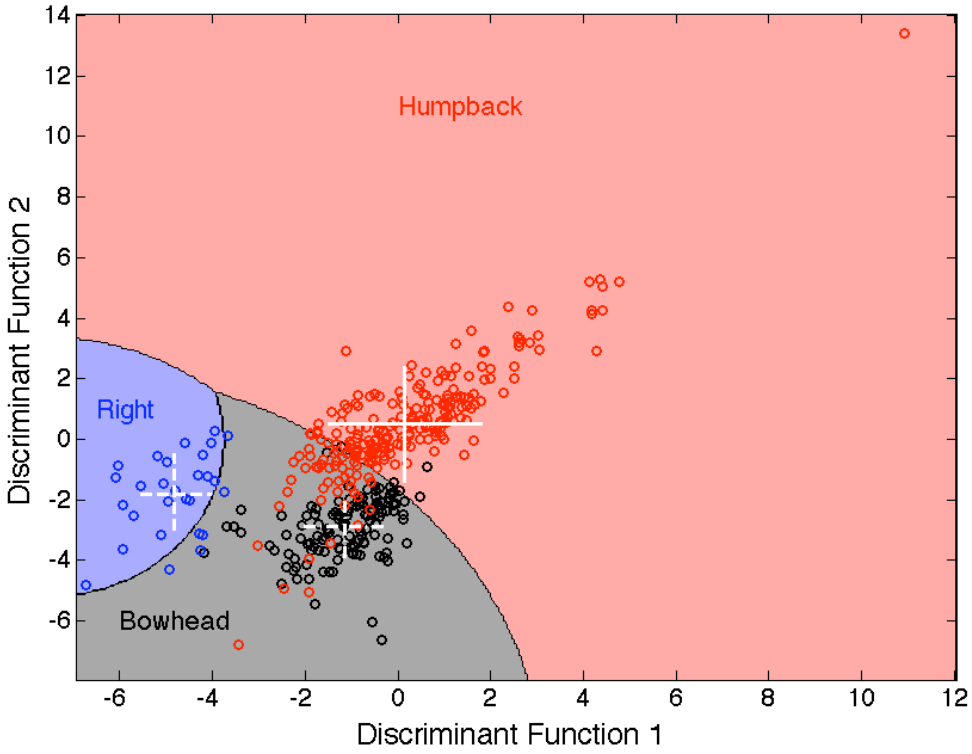


Figure 7.5 Decision region for classification of bowhead, humpback and right whale vocalizations. Data points from the testing subset were projected onto the 2D space using DA. When three classes are used, DA produces only the two discriminant functions used for plotting. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

Table 7.7 Confusion matrix of *AUC* values corresponding to the decision region shown in Figure 7.5, where feature space dimensionality reduction is performed using DA. The value $M = 0.98$. The asterisk indicates *AUC* values that appear to result from an ideal classifier, but in fact resulted from rounding to two decimal places.

| | **Humpback** | **Right** |
|---|---|---|
| **Bowhead** | 0.93 | 1.00 |
| **Humpback** | | 1.00* |

By visually examining the decision regions in Figure 7.4 and Figure 7.5, it is apparent that, in general, the data points are clustered more tightly around their respective class means for the DA case compared to the PCA case. Distance between class means increased when DA replaced PCA, as can be noted from the summary statistics reproduced in Table 7.8. There were no statistically significant improvements in results ($\Delta M = 0.0004$) gained by implementing DA in the three-class case analyzed here. The decrease in overall classification accuracy occurred because of an increase in the number of misclassified humpback vocalizations; however, more right whale vocalizations were correctly classified – this is an example of class skew negatively affecting the accuracy performance measure.

Table 7.8 Summary statistics describing the distance between class means of baleen whale data points when reduced feature spaces are composed of either principal components or discriminant functions. The three class means correspond to the white crosses displayed on Figure 7.4 and Figure 7.5.

| | **2 Principal Components** | **2 Discriminant Functions** |
|---|---|---|
| **Minimum distance** | 2.96 | 3.61 |
| **Maximum distance** | 4.93 | 5.51 |
| **Mean distance** | 4.13 | 4.33 |

Eigenvectors used for projecting data points onto the lower-dimensional feature space using either PCA or DA are represented in Figure 7.6. The three highest weighted features generated by PCA and DA are listed in Table 7.9. Similar relative weightings are not applied to the features used in PCA and DA (i.e. when plotted as shown below, the sums of the eigenvectors do not exhibit a similar trend). Only peak loudness value is

included in the top three features with the most influence in both the principal components and discriminant functions.

Table 7.9 Three highest weighted features using PCA and DA methods.

| **PCA Features** | **DA Features** |
|---|---|
| Psychoacoustic maxima-to-spectral bins ratio | Peak loudness value |
| Global mean sub-band attack time | Mean sub-band correlation |
| Peak loudness value | Integrated loudness |

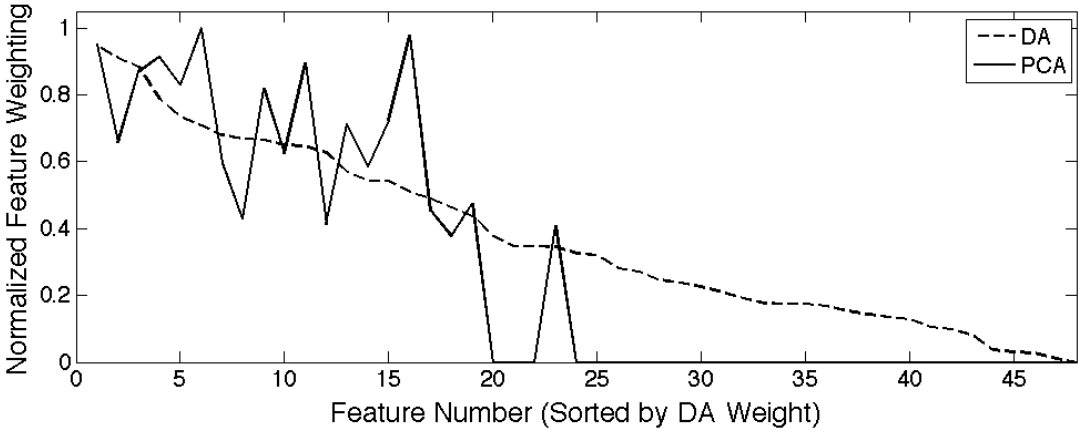


Figure 7.6 Normalized weighting of features for classification when using either PCA or DA for dimensionality reduction during classification of baleen whales. Features are sorted from largest DA feature weighting to smallest. These eigenvectors correspond to the PCA-based decision region shown in Figure 7.4 and DA-based decision region in Figure 7.5. Peaks are connected merely for visualization purposes and are not intended to imply that the data are continuous.

### 7.1.3 Two classes ($c$ = 2)

Discriminant analysis with two classes produces a single discriminant function. Because there is only one discriminant function, the 2D decision regions shown in previous sections cannot be generated. Instead, the data is plotted in one-dimension with a single (i.e. one-dimensional) threshold value deciding to which class each sample belongs. As stated previously, the only limit on the number of principal components produced by PCA is the number of dimensions being considered, i.e. if twenty features are selected for classification, PCA can produce up to twenty principal components.

The decision region generated using PCA for classification of bowheads and humpbacks is shown in Figure 7.7. The accuracy of classification was 87%. The corresponding ROC curve is shown in Figure 7.9 with $AUC = 0.95$ and an equal error rate of 12%. There is more scatter evident in the humpback class than the bowhead class because of the greater variety of calls used.

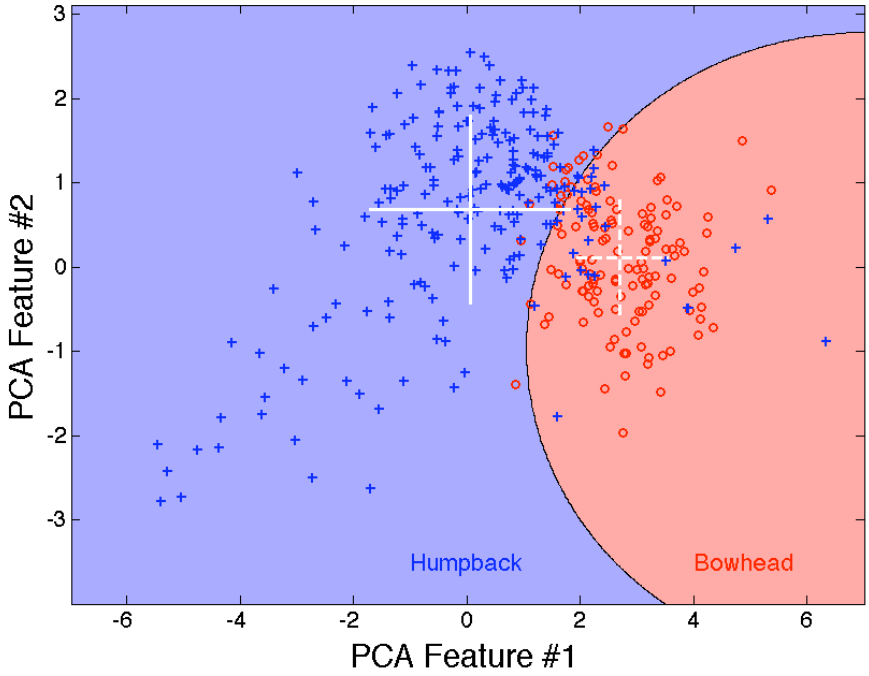


Figure 7.7 Decision region for classification of bowhead and humpback vocalizations. Data points from the testing subset were projected onto the 2D space using PCA with twenty selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

Classification results for bowhead and humpback whales, produced using DA for feature space reduction, are shown in Figure 7.8. Results are summarized using histograms for each whale class; a total of 45 bins were used to generate the histograms. As is the case for the 2D decision regions used to represent previous results, background colour is used to represent correct or incorrect classification decisions; for example, if the light-coloured bar representing binned bowhead vocalizations is located on the white background the vocalizations contained in that bin were correctly classified, whereas if a light-coloured bar is located on the grey background an incorrect classification decision was made. As

can be seen from the decision region, more humpback vocalizations were misclassified as bowheads than bowheads misclassified as humpbacks – only three bowhead vocalizations were misclassified. The shape of the humpback distribution is wider, indicating greater variance in the humpback class. This is as expected given the broad range of humpback vocalizations. The accuracy of classification was 88%. The ROC curve corresponding to this classification is represented in Figure 7.9, which has an *AUC* of 0.96 and an equal error rate of 9%.

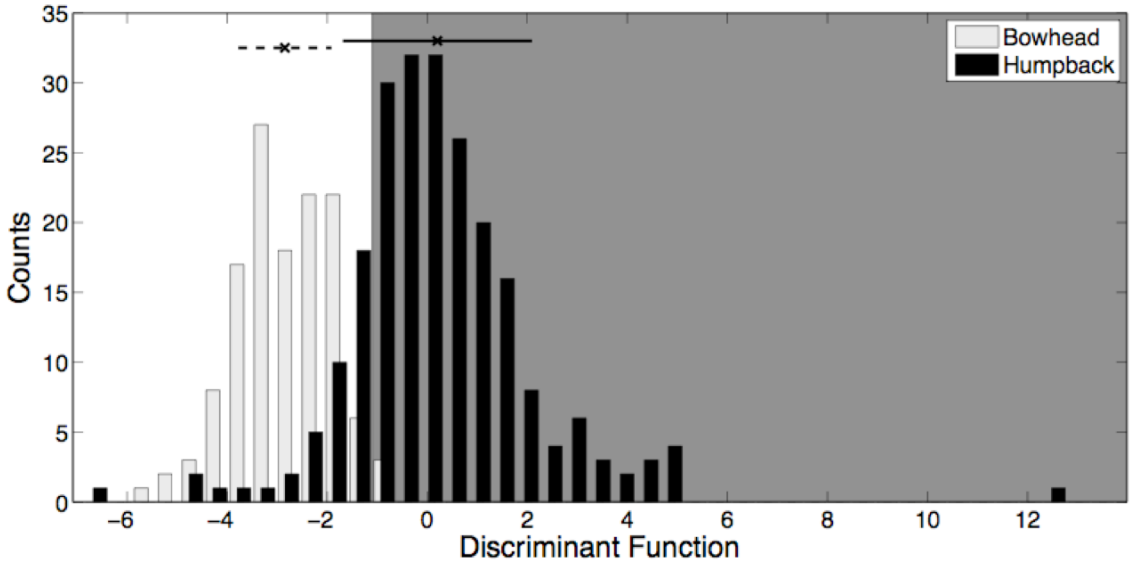


Figure 7.8 Histogram representing bowhead versus humpback classification results. Discriminant analysis was used to perform feature space reduction. Background colouring represents the classification decision, for example all black bars that fall in the grey region represent correctly classified vocalizations and any black bars that fall in the white region correspond to incorrectly classified vocalizations. The two horizontal lines above the histograms have length of one standard deviation from their respective means (represented by the crosses). The dashed line corresponds to the bowhead data and the solid line to the humpback distribution.

As a side note, visually the histograms shown in the DA-based decision region (Figure 7.8) may resemble a one-dimensional Gaussian PDF. The null hypothesis that both the bowhead and humpback classes are Gaussian distributed along the discriminant function was tested using the $\chi^2$ goodness of fit test to quantify the resemblance. It was found that the null hypothesis could not be rejected for the bowhead class but is rejected for the humpback class. This is because the chances that a random sample from a Gaussian

distribution would produce a value equal to or larger than the calculated $\chi^2$ test statistic are 37% and less than 1% for the bowhead and humpback classes, respectively. A 50% probability indicates an ideal fit because statistically observed values of $\chi^2$ should only exceed the norm half the time [54]. Thus, the results of the $\chi^2$ goodness of fit test provide further evidence as to the validity of using a Gaussian-based classifier.

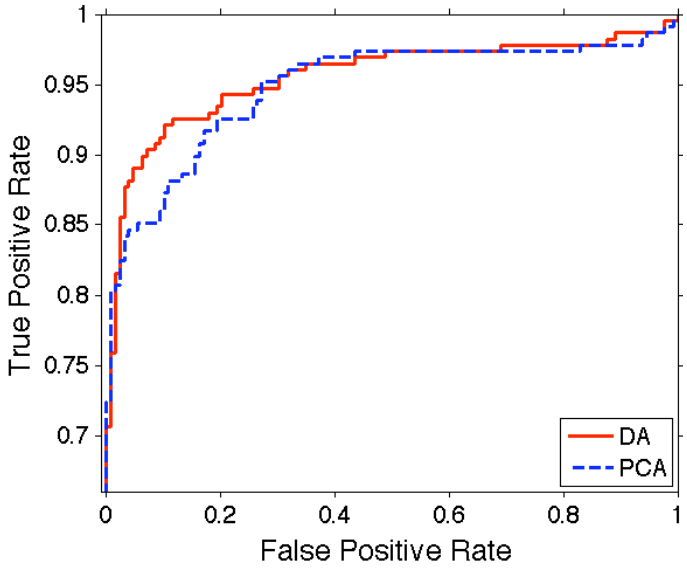


Figure 7.9 ROC curves resulting from classification of bowheads and humpbacks using either PCA or DA for feature space reduction. Only the region where the ROC curves do not overlap is plotted. These curves correspond to the decision regions shown in Figure 7.7 and Figure 7.8 . When using PCA, $AUC = 0.95$ and when using DA, $AUC = 0.96$.

Small improvements in classification results were obtained by implementing DA for classification of bowhead and humpback vocalizations. PCA produced a transformation that caused a similar number of bowhead and humpback vocalizations to be misclassified; whereas when the DA method was used the ratio of misclassified humpbacks to bowheads was noticeably large. Implementing DA resulted in only a one percentage point increase in accuracy. The $AUC$ value using DA was also larger than the value obtained using PCA ($\Delta AUC = 0.01$). The distance between class means are summarized in Table 7.11 – as expected the distance between class means is larger when the DA method is used. The feature weighting, as determined by the sum of eigenvectors, is represented in Figure 7.10. The three features with the greatest

weighting using PCA and DA are listed in Table 7.10 and relative feature weighting is represented in Figure 7.10. There are no commonalities in features with large weightings when using either PCA or DA.

Table 7.10 Three highest weighted features using PCA and DA methods.

| **PCA Features** | **DA Features** |
|---|---|
| Local mean sub-band decay slope | Peak loudness value |
| Global mean sub-band attack time | Global maximum sub-band decay time |
| Global mean sub-band decay slope | Mean sub-band correlation |

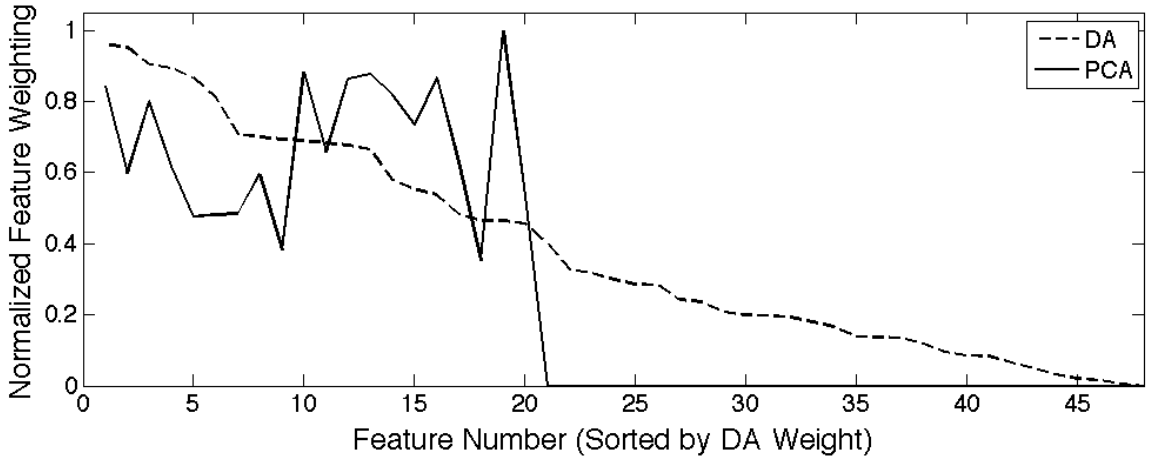


Figure 7.10 Normalized weighting of features used for classification when using either PCA or DA for dimensionality reduction during classification of bowhead and humpback whales. Features are sorted from largest DA feature weighting to smallest. These eigenvectors correspond to the PCA-based decision region shown in Figure 7.7 and DA-based decision region in Figure 7.8. Peaks are connected merely for visualization purposes and are not intended to imply that the data are continuous.

Classification of bowhead and right whale 1&2 vocalizations was also performed to determine differences in classification results between PCA and DA when only two classes are considered. The decision region, whose axes are the first two principal components, is shown in Figure 7.11. PCA performs well with these two classes as can be noted by the relatively large separation of class means (white crosses shown in the figure) and the relatively small number of misclassifications (accuracy of 97%). The ROC curve corresponding to this classification is represented in Figure 7.13. The $AUC = 1.00$, or nearly ideal, and a relatively low equal error rate of 4% which is also indicative of a classifier that is near ideal.

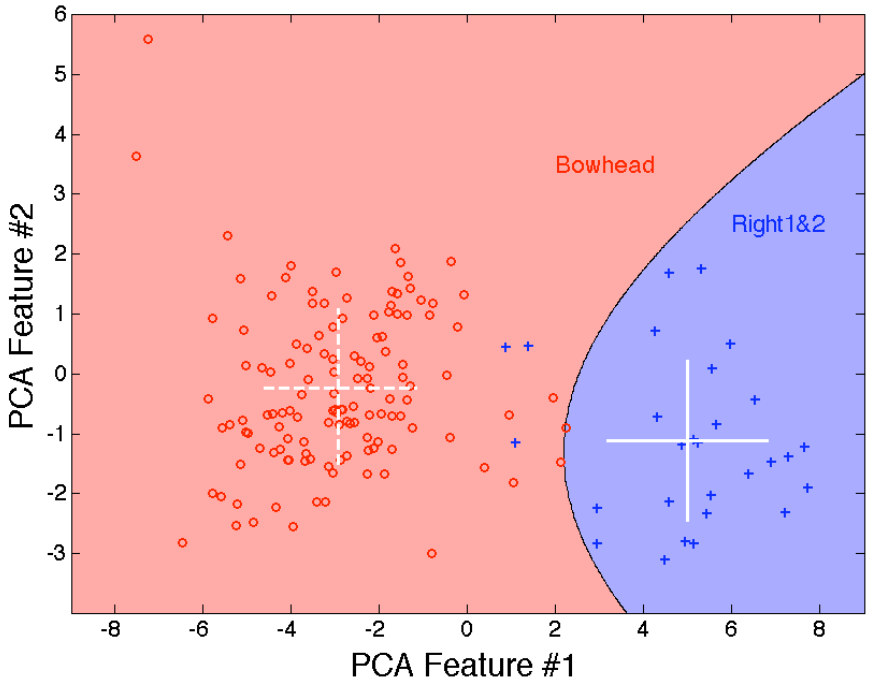


Figure 7.11 Decision region for classification of bowhead and right whale1/right whale2 vocalizations. Data points from the testing subset were projected onto the 2D space using PCA with twenty selected features. Class means are represented as white crosses on their respective decision regions with bars one standard deviation in length.

Classification results for bowhead and right whale 1&2 vocalizations using DA for dimensionality reduction are represented by the histogram decision region (Figure 7.12), which shows a 98% accurate classification. The ROC curve in Figure 7.13 has an $AUC = 1.00$ and a low equal error rate of 4%. Good separation of class means is apparent by examining the decision region and observing that there is little overlap between the two classes. Only three right whale vocalizations were misclassified. Examining the shape of the right whale vocalization histogram, there appears to be relatively high scatter in this class with an approximately uniform distribution of data points along the discriminant function compared to the bowhead class which exhibits a possible Gaussian shape (i.e. there are more data points distributed about the class mean and smooth tapering off on either side of the mean). The $\chi^2$ goodness of fit test concludes that the null hypothesis, which assumes the data in each class are Gaussian distributed, is rejected in the case of the right whales but is not rejected for the bowhead distribution because there is an 88% and a 64% chance that a random sample of data from a Gaussian would

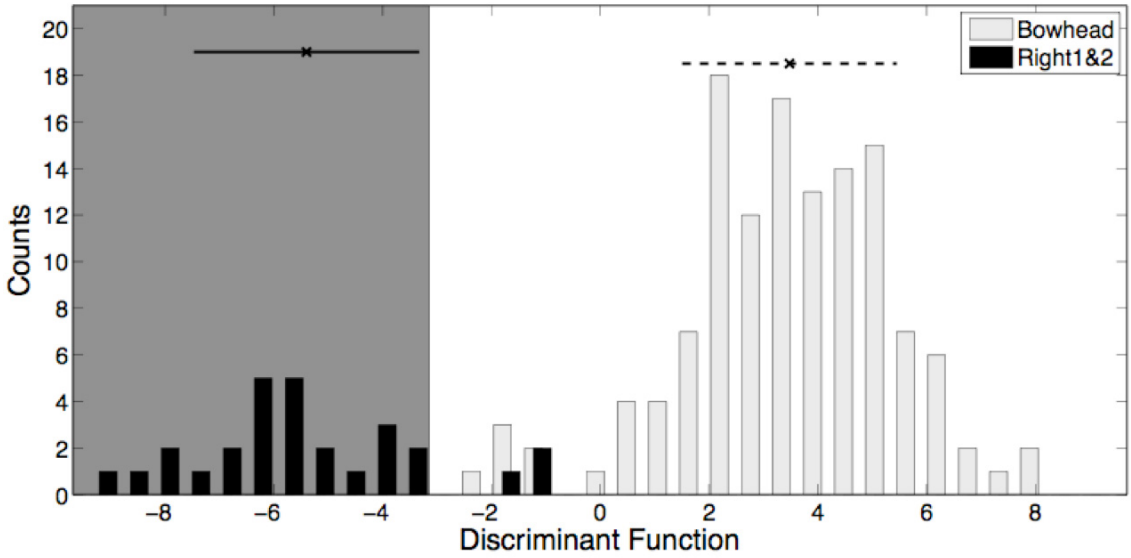


Figure 7.12 Histogram representing bowhead versus North Atlantic right whale 1&2 classification results. Discriminant analysis was used to perform feature space reduction. Background colouring represents the classification decision, black bars that fall in the grey region represent correctly classified vocalizations and any black bars that fall in the white region correspond to incorrectly classified vocalizations. The two horizontal lines above the histograms have length of one standard deviation from their respective means (represented by the crosses). The dashed line corresponds to the bowhead data and the solid line to the right whale distribution.

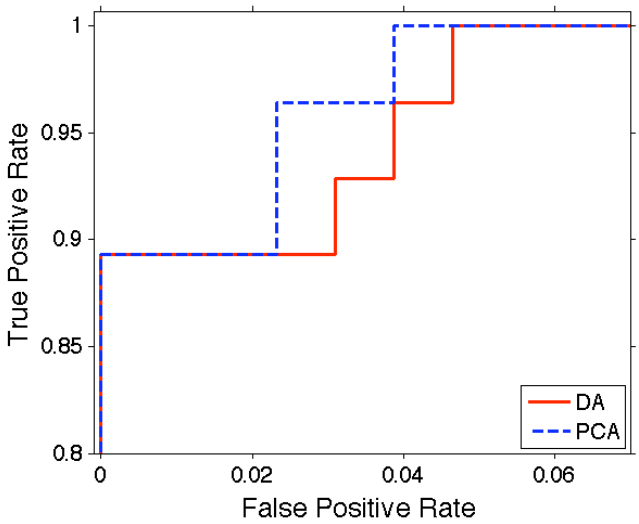


Figure 7.13 ROC curves from classification of bowhead and right1&2 using either PCA or for feature space reduction. Only the region where the ROC curves do not overlap is plotted. These curves correspond to decisions regions in Figure 7.11 and Figure 7.12. When using PCA, $AUC = 1.00$ and when using DA, $AUC = 1.00$.

produce a larger $\chi^2$ value than given for the right whale and bowhead distributions, respectively. This may be due to the small size of the right whale dataset – not enough vocalizations were used to adequately represent the distribution of this class with respect to the given discriminant function.

Classification results for bowhead and right whale1&2 were similar between the PCA and DA methods. There was no significant difference between the ROC curves and *AUC* values. The same number of right whale vocalizations was misclassified in each case, although one bowhead vocalization was misclassified in the PCA method, whereas no bowhead vocalizations were misclassified using the discriminant function. Comparing the distance between the class means computed for each of the feature space reduction methods (summarized in Table 7.11) shows that there was greater separation between the class means when the PCA method was used. This is unlike the previously discussed examples where the distance between means was greater when DA was implemented. The PCA method may produce a greater distance between the means because two dimensions were used, compared to the DA case where it was possible to produce only a single discriminant function. This unexpected result may also be due to the small size of the right whale dataset causing the distribution to be undersampled.

Table 7.11 Distance between class means for examples of $c = 2$ when reduced feature spaces are composed of either principal components or discriminant functions. For the PCA cases, class means correspond to the white crosses displayed on Figure 7.7 and Figure 7.11.

| | **PCA** | **DA** |
|---|---|---|
| **Bowhead/ Humpback** | 2.08 | 5.83 |
| **Bowhead/ Right 1&2** | 7.06 | 6.81 |

Feature weightings are represented graphically in Figure 7.14. The three features with the highest weighting as determined by PCA and DA are listed in Table 7.12. Global maximum sub-band decay time was the only feature to receive a large weight value from both DA and PCA methods. However, generally, there does not appear to be much similarity in the relative feature weightings between the PCA and DA methods indicating that the features that best represent the variation in the dataset do not correspond to the features that best separate the bowhead and right whale vocalization classes.

Table 7.12 Three highest weighted features using PCA and DA methods.

| PCA Features | DA Features |
|---|---|
| Local maximum sub-band decay time | Global maximum sub-band decay time |
| Global maximum sub-band decay time | Pre-attack psychoacoustic maxima to spectral bins ratio |
| Local maximum sub-band attack time | Integrated loudness |

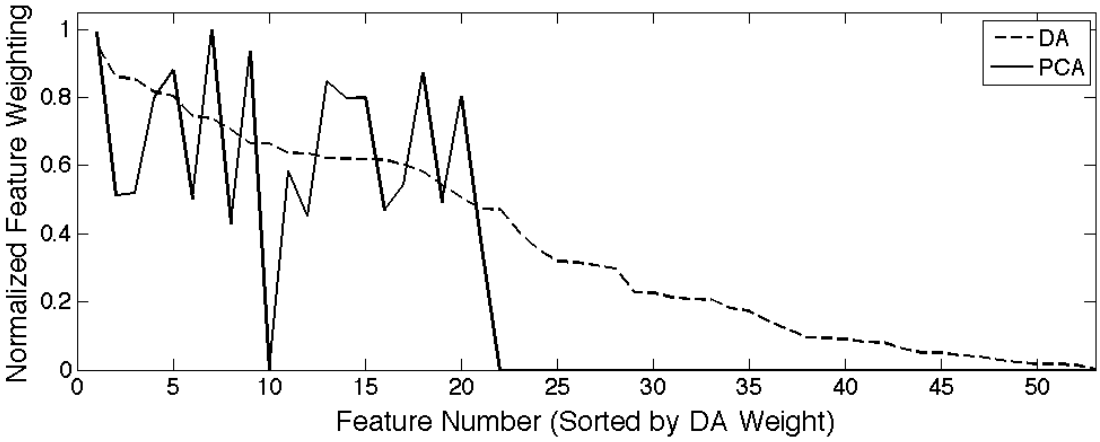


Figure 7.14 Normalized weighting of features for classification when using either PCA or DA for dimensionality reduction during classification of bowhead and North Atlantic right 1 and 2 vocalizations. Features are sorted from largest DA feature weighting to smallest. These eigenvectors correspond to the PCA-based decision region shown in Figure 7.11 and DA-based decision region in Figure 7.12. Peaks are connected merely for visualization purposes and are not intended to imply that the data are continuous.

## 7.2 CONCLUSIONS

Implementing discriminant analysis resulted in improvements to aural classification results; in fact, all examples showed improvements in results except for the binary bowhead and right whale classification example. The small size of the right whale dataset was likely the main factor in decreased performance when DA was used. Improvements gained by implementing DA were particularly evident for automatic classification using larger numbers of classes. When more than two discriminant functions could be computed it was found that classification results were better when all discriminant functions were used, because it allowed for a greater separation of class means; however in all of the cases examined, classification performance in the PCA-based feature space was already very good so that only slight improvements were

realistically possible. Only slight improvements in results when using PCA or DA may have resulted because the features combined in the principal components were selected based on good between-class discriminability (i.e. using the Fisher Linear Discriminant Score).

For $c = 2$ there were no significant changes in classifier effectiveness noted by implementing DA. The most obvious disadvantage of DA when two classes are examined is that only a single discriminant function can be produced – the ability to use two or more principal components lends an advantage to the use of PCA because it provides the opportunity for greater separation of class means by adding an extra degree of freedom. However, classification results were not diminished by implementing DA, thus, this method can be used in place of PCA with the expectation of similar classification results.

An indirect advantage of DA is that it provides an upper limit on the number of discriminant functions, based on the number of classes under consideration, and thus the dimensionality of the transformed feature space. PCA, on the other hand, provides as many eigenvectors as features being considered – it is then up to the researcher to decide the number of principal components to use, which may be somewhat arbitrary. Additionally, this implementation of DA does not require pre-selecting the best features for classification as is done with the PCA method, but instead generates the projection onto the lower dimensional space by considering all non-redundant features. This removes a processing step and the (somewhat) arbitrary selection of the number of features to include in PCA that was shown to be a non-trivial problem in Section 5.3.1.

Computationally, implementing DA caused no significant changes; no noticeable change in run time for the automatic classifier was noted by implementing DA. The implementation of discriminant analysis relies on computing the inverse of the within-class scatter matrix – if this matrix is not independent, and therefore non-invertible, then it will not be possible to compute the discriminant functions. However, it is unlikely that

the within-class scatter will not be independent, so this is only a minor concern when using DA for transformation of the feature space.

The features dominating the choice of discriminant functions were typically different than those with high weighting in the principal components. In most of the examples, there was more variation in the relative weightings used by the DA eigenvectors, whereas there was less variation in the weightings of features used in PCA. This is likely due to the fact that the number of features considered was narrowed down prior to performing PCA. Based on the relative weighting of features in the principal components and discriminant functions it is possible to conclude that the features that best separate the vocalization classes do not necessarily correspond to the features that best represent the variance in the dataset.

Discriminant analysis has been shown to be a useful addition to the automatic aural classifier by improving classification results in the four different examples discussed. The benefits of using DA are more noticeable when three or more classes are used. Replacing PCA with DA and testing results on the cetacean data verified the trends predicted by theory – there was a noticeable decline in the within class scatter and a greater separation of class means. It is recommended that DA continue to be employed by the automatic aural classifier in the future.

# CHAPTER 8 SUMMARY AND CONCLUSIONS

## 8.1 SUMMARY AND CONCLUSIONS

The aural classifier has proven to be a useful tool for classifying marine mammal vocalizations. Multiclass classification with the five cetacean species in the dataset (bowhead, humpback, North Atlantic right, minke, and sperm whales) produced accurate results; the best results were obtained when five features were selected to include in the principal components. The overall accuracy of classification was 89% and the multiclass performance measure was near ideal ($M = 0.99$). The aural classifier easily distinguished between sperm whale clicks and baleen whale vocalizations with high accuracy using both multiclass and binary classification methods. Binary classification of pairs of whale species resulted in improved classifier performance on baleen whales because features were selected and weighted for discrimination between only two species, rather than for recognizing the patterns in all five classes. The baleen whale binary classification results were all near ideal. The most challenging binary case was for classification of bowhead and humpback vocalizations because of both the signal (i.e. before auditory model is applied) and aural similarities of these species' vocalizations, and due to the large variety of sounds made by humpbacks; however, the aural classifier was able to successfully discriminate between the vocalizations of bowhead and humpback whales with 92% accuracy, and achieved an $AUC = 0.97$ and equal error rate of 5%.

Sperm whale clicks and a variety of anthropogenic passive transients were successfully discriminated using the aural classifier. Classification performance was characterized by 98% accuracy, an $AUC$ of 1.00 and equal error rate of 1%; each of these three

performance metrics are indicative of near-ideal classification. When classification of sperm whale clicks and anthropogenic transients was performed, it was noted that the data points were arranged in a strongly correlated linear pattern within their corresponding classes. This linear within-class trend was also observed in the sperm whale class during multiclass cetacean classification. Analysis of the multiclass cetacean classification and the binary anthropogenic transients/sperm whale results indicated that the linear trend in the PCA space occurred when highly ranked features (or feature combinations) had a relatively small amount of variance along the semi-minor axis of the within-class scatter and a relatively large amount of variance along the semi-major axis, due to a correlation between features.

Classification of cetacean vocalizations primarily relied on purely spectral perceptual features. This indicated there were measurable differences in the perceptual spectra of the vocalizations that could be taken advantage of for between-class discrimination. Conversely, time-frequency perceptual features were highly ranked for classification of sperm whale clicks and anthropogenic passive transients. Time-frequency features successfully described the differences between the implosive non-reverberant characteristics of sperm whale clicks and the slower attack and decay times of the anthropogenic transients.

This research also provided the opportunity to enhance the aural classifier in two ways. First, multiclass performance measures were implemented to provide a method for analyzing the effectiveness of the classifier similar to the *AUC* metric used in the binary case. The *M*-measure and corresponding confusion matrix of pairwise *AUC* values were shown to be useful metrics for qualitatively evaluating multiclass classification results. Second, discriminant analysis was implemented. Multiclass classification results were improved by replacing PCA with DA mainly because distance between class means increased. However, for binary classification no significant performance improvements were observed when DA replaced PCA – this was because discriminant analysis was used to select features to include in the principal components and DA with two classes produces only a single discriminant function.

The aural classifier performed well when qualitatively compared to results from other automatic detection and/or classification algorithms. Many of the marine mammal detection and classification algorithms presented in the literature are based on correlation techniques and so are often specific to a single type of vocalization; the aural classifier uses the distinct aural features of species' vocalizations to inform classification decisions and can therefore simultaneously classify vocalizations from multiple species. This showed the aural classifier to be more robust than many of the methods presented in the literature.

## 8.2 SUGGESTIONS FOR FUTURE WORK

Aural classification of cetacean vocalizations displayed the effectiveness of the classifier for distinguishing between marine mammal species based on their vocalizations. The results were very promising and pave the way for additional future work.

Future work should focus on testing the robustness of the aural classifier to progress towards the ambitious goal of automatically performing aural classification in real-time or near-real-time scenarios. The dataset used in this research was relatively limited – in order to gain a better understanding of how the aural classifier deals with within-class variance the size of the dataset should be increased. Marine mammal vocalizations often vary from individual-to-individual, by geographic region, by season, and by behaviour. To gain a better understanding of how the aural classifier would perform in a realistic scenario, both the marine mammal and anthropogenic transient datasets should be supplemented with data that captures more within-species (or within-class) variation. Contextual information may also be included to possibly enhance classifier performance. The seasonality of marine mammal sightings may be incorporated in the system since many whales are known to migrate according to time of year; this knowledge can be used to weight the likelihood probabilities. For example, humpbacks are not known to stay in the Bay of Fundy during the winter months so if the aural classifier detected a possible humpback whale vocalization, the likelihood probability, $P(H|\mathbf{x}')$, would need to be high in order to identify the vocalization as resulting from a humpback.

The perceptual features used in this research were originally selected to discriminate active sonar echoes. It may be beneficial to investigate additional perceptual features that may better discriminate between marine mammal vocalizations. Additionally, incorporating features that describe spectrogram characteristics may better represent the way in which human analysts identify the species vocalizing, since human experts use a combination of aural cues and visual inspection of spectrograms to identify species.

# BIBLIOGRAPHY


[1] Alexandros Frantzis, "Does acoustic testing strand whales?," *Nature*, vol. 392, p. 29, March 1998.

[2] Lindy S. Weilgart, "The impacts of anthropogenic ocean noise on cetaceans and implications for management," *Canadian Journal of Zoology*, vol. 85, pp. 1091-1116, 2007.

[3] Peter L. Tyack et al., "Beaked whales respond to simulated and actual navy sonar," *PLoS ONE*, vol. 6, no. 3, p. e17009, March 2011.

[4] Xavier Mouy, Del Leary, Bruce Martin, and Marjo Laurinolli, "A comparison of methods for automatic classification of marine mammal vocalizations in the Arctic," in *New Trends for Environmental Monitoring Using Passive Acoustic Systems, 2008*, Hyeres, 2008, pp. 1-6.

[5] Denise Risch, Peter J. Corkeron, William T. Ellison, and Sofie M. Van Parijs, "Changes in humpback whale song in response to an acoustic source 200 km away," *PLoS ONE*, vol. 7, no. 1, p. e29741, January 2012.

[6] Right wale listening network: bioacoustics research program. [Online]. www.listenforwhales.org/

[7] Angelia S.M. Vanderlaan, Christopher T. Taggart, Anna R. Serdynska, Ronald D. Kenney, and Moira W. Brown, "Reducing the risk of lethal encounters: vessels and right whales in the Bay of Fundy and on the Scotian Shelf," *Endangered Species Research*, vol. 4, pp. 283-297, April 2008.

[8] "Defence S&T Strategy: Science and Technology for a Secure Canada," Department of National Defence, 2006.

[9] Victor W. Young and Paul C. Hines, "Perception-based automatic classification of impulsive-source active sonar echoes," *Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1502-1517, September 2007.

[10] Whitlow W.L. Au and Mardi C. Hastings, *Principles of Marine Bioacoustics*, 1st ed., Robert T. Beyer and William Hartmann, Eds. New York, USA: Springer Science, 2008.

[11] David K. Mellinger, Kathleen M. Stafford, Sue E. Moore, Robert P. Dziak, and Haru Matsumoto, "An overview of fixed passive acoustic observation methods for cetaceans," *Oceanography*, vol. 20, no. 4, pp. 36-45, December 2007.

[12] David K. Mellinger, "A comparison of methods for detecting right whale calls," *Canadian Acoustics*, vol. 32, no. 2, pp. 55-65, 2004.

[13] Brian R. La Cour and Michael A. Linford, "Detection and classification of North Atlantic right whales in the Bay of Fundy using independent component analysis," *Canadian Acoustics*, vol. 32, no. 2, pp. 48-54, 2004.

[14] A. T. Johansson and P. R. White, "Detection and characterization of marine mammal calls by parametric modelling," *Canadian Acoustics*, vol. 32, no. 2, pp. 83-92, 2004.

[15] A. Moscrop, J. Matthews, D. Gillespie, and R. Leaper, "Development of passive acoustic monitoring systems for northern right whales," *Canadian Acoustics*, vol. 32, no. 2, pp. 17-22, 2004.

[16] Peter J. Dugan, Aaron N. Rice, Ildar R. Urazghildiiev, and Christopher W. Clark, "North Atlantic right whale acoustic signal processing: Part I comparison of machine learning recognition algorithms," in *IEEE Applications and Technology Conference (LISAT), 2010 Long Island Systems*, Farmingdale, NY, 2010, pp. 1-6.

[17] Cynthia T. Tynan and Douglas P. DeMaster, "Observations and predictions of Arctic climatic change: potential effects on marine mammals," *Arctic*, vol. 50, no. 4, pp. 308-322, December 1997.

[18] William A. Yost, *Fundamentals of Hearing: An Introduction*, 4th ed. San Diego, California, USA: Academic Press, 2000.

[19] Nancy Allen, Paul C. Hines, and Victor W. Young, "Performances of human listeners and an automatic aural classifier in discriminating between sonar target echoes and clutter," *Journal of the Acoustical Society of America*, vol. 130, no. 3, pp. 1287-1298, September 2011.

[20] W. Victor Young, "Application of musical timbre discrimination features to active sonar classification," Physics and Atmospheric Science, Dalhouise University, Halifax, NS, M.Sc. Thesis 2005.

[21] Jorg Kliewer and Alfred Mertins, "Audio subband coding with improved representation of transient signals," in *Proceedings of the IX European Signal Processing Conference*, Rhodes, Greece, 1998, pp. 1245-1248.

[22] Malcolm Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Perception Group - Advanced Technology Group, Apple Computer Inc., Technical TR-35, 1993.

[23] Julius O. Smith and Jonathan S. Abel. (2007) Center for Computer Research in Music and Acoustics, Stanford University. [Online]. https://ccrma.stanford.edu/~jos/bbt/Equivalent_Rectangular_Bandwidth.html

[24] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, 2nd ed. Toronto, ON, Canada: Wiley-Interscience, 2001.

[25] Tom Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861-874, 2006.

[26] David J. Hand and Robert J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Machine Learning*, vol. 45, pp. 171-186, 2001.

[27] Daivd M. Green and John A. Swets, "Chapter 2: Statistical decision theory and psychophysical procedures," in *Signal Detection Theory and Psychoacoustics*. Los Altos, California, USA: Peninsula Publishing, 1988, pp. 30-52.

[28] Mark F. Davis, "The psychoacoustics of audio and electroacoustics," in *Springer Handbook of Acoustics*, Thomas D. Rossing, Ed. Ney York, USA: Springer Science, 2007, p. 747.

[29] Julie N. Oswald, Whitlow W.L. Au, and Fred Duennebier, "Minke whale (Balaenoptera acutorostrata) boings detected at the Station ALOHA Cabled Observatory," *Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3353-3360, May 2011.

[30] Government of Canada - Environment Canada. (2011, September) Species at Risk Public Registry. [Online]. http://www.sararegistry.gc.ca/default_e.cfm

[31] Joe D. Hood and Derek Burnett, "Compilation of marine mammal passive transients for aural classification," Defence R&D Canada - Atlantic, Halifax, NS, Contract Report CR 2008-287, 2009.

[32] Sara Heimlich, Holger Klinck, and Dave Mellinger. (2010) MobySound. [Online]. http://www.mobysound.org

[33] David K. Mellinger and Christopher W. Clark, "MobySound: A reference archive for studying automatic recognition of marine mammal sounds," *Applied Acoustics*, vol. 67, pp. 1226-1242, 2006.

[34] (2008) Home. [Online]. www.jasco.com/

[35] Roger S. Payne and Scott McVay, "Songs of humpback whales," *Science*, vol. 173, pp. 585-597, August 1971.

[36] Alyson Fleming and Jennifer Jackson, "Global review of humpback whales (Megaptera novaeangliae)," National Marine Fisheries Service, Southwest Fisheries Science Center, National Oceanic and Atmospheric Administration, Technical Memorandum NOAA-TM-NMFS-SWFSC-474, 2011.

[37] H. E. Winn et al., "Song of the humpback whale - population comparisons," *Behavioural Ecology and Sociobiology*, vol. 8, pp. 41-46, 1981.

[38] Alison K. Stimpert, Whitlow W.L. Au, Susan E. Parks, Thomas Hurst, and David N. Wiley, "Common humpback whale (Megaptera novaeangliae) sound types for passive acoustic monitoring," *Journal of the Acoustical Society of America*, vol. 129, no. 1, pp. 476-482, January 2011.

[39] Science Daily. (2005, July) Science Daily. [Online]. http://www.sciencedaily.com/releases/2005/07/050726075715.htm

[40] Bjarke Klit Nielsen and Bertel Mohl, "Hull-mounted hydrophones for passive acoustic detection and tracking of sperm whales (Physete macrocephalus)," *Applied Acoustics*, vol. 67, pp. 1175-1186, 2006.

[41] Aaron Thode, David K. Mellinger, Sarah Stienessen, Anthony Martinez, and Keith Mullin, "Depth-dependent acoustic features of diving sperm whales (Physeter macrocephalus) in the Gulf of Mexico," *Journal of the Acoustical Society of Amemrica*, vol. 112, no. 1, pp. 308-321, July 2002.

[42] Shannon Rankin and Jay Barlow, "Source of the North Pacific "boing" sound attributed to minke whales," *Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3346-3351, November 2005.

[43] Cooperative Institute for Marine Resources Studies. (2010) CIMRS Bioacoustics Lab. [Online]. http://www.bioacoustics.us/dcl.html

[44] (2011, July) SoX - Sound eXchange. [Online]. http://sox.sourceforge.net

[45] Joe D. Hood and David G. Flogeras, Improved passive band-limited energy detection for marine mammals, May 3, 2011, Private communication.

[46] David K. Mellinger and Christopher W. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation," *Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3518-3529, June 2000.

[47] Douglas Gillespie, "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," *Canadian Acoustics*, vol. 32, no. 2, pp. 39-47, 2004.

[48] Mark F. Baumgartner and Sarah E. Mussoline, "A generalized baleen whale call detection and classification system," *Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 2889-2902, May 2011.

[49] Joe D. Hood and Ben Bougher, "Compilation of Anthropogenic Passive Transients for Aural Classification," Defence R&D Canada - Atlantic, Dartmouth, NS, Contractor Report.

[50] Gary Nichols, "Chapter 22: Subsurface stratigraphy and sedimentology," in *Sedimentology and Stratigraphy*, 2nd ed. West Sussex, United Kingdom: Wiley-Blackwell, 2009, pp. 335-348.

[51] Mary L. Boas, *Mathematical Methods in the Physical Sciences*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons Inc., 2006.

[52] V. Kandia and Y. Stylianou, "Detection of sperm whale clicks based on the Teager-Kaiser energy operator," *Applied Acoustics*, vol. 67, pp. 1144-1163, 2006.

[53] Quyen Huynh, Walter Greene, and John Impagliazzo, "Feature extraction and classification of underwater acoustic signals," in *Full Field Inversion Methods in Ocean and Seismo-Acoustics*, O. Diachok et al., Eds. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1995, pp. 183-188.

[54] Philip R. Bevington and D. Keith Robinson, "4.3 Chi square test of a distribution ," in *Data Reduction and Error Analysis for the Physical Sciences*, 2nd ed. New York, USA: McGraw-Hill, 1992, pp. 65-70.

# APPENDIX A PERCEPTUAL FEATURES

The following table contains a summary of all the perceptual signal features, separated into time-frequency and purely spectral perceptual signal features. Qualitative descriptions of how each of the features is calculated are also included (as in [9]).

| | **Perceptual Signal Feature** | **Quantitative Representation** | **Description** |
|---|---|---|---|
| **Time-frequency** | Sub-band attack | Time (SBAT) | Time delay between the KM-defined vocalization start and the peak of the temporal envelope |
| | | Slope (SBAS) | Slope of the line joining the start of the vocalization and the peak of the temporal envelope |
| | Sub-band decay | Time (SBDT) | Time delay between the peak of the temporal envelope and the KM-defined vocalization end |
| | | Slope (SBDS) | Slope of the line joining the peak of the temporal envelope and the end of the vocalization |
| | Sub-band synchronicity | Correlation (SBCorr) | Average correlation coefficient between the temporal envelope for the $i^{th}$ channel and the temporal envelopes for the remaining channels: $\frac{1}{99}\sum_{j \neq i} \rho_{i,j}$ |
| **Spectral** | Peak loudness | Peak loudness frequency (PLF) | Centre frequency (in ERB) of the filter bank channel containing the maximum value of the perceptual loudness function |
| | | Peak loudness value (PLV) | Value of the perceptual loudness function (in sones/ERB) corresponding to the PLF |

| | Perceptual Signal Feature | Quantitative Representation | Description |
|---|---|---|---|
| | Loudness Roughness | Maxima to spectral bin ratio (MSBR) | The total number of local maxima in the perceptual loudness function divided by the number of filter bank channels |
| | | Bin-to-bin difference (BBD) | The mean of the magnitude of the difference between adjacent bins (i.e. filter bank channels) of the perceptual loudness spectrum |
| | Loudness centroid | Loudness centroid (LC) | The frequency (in ERB) corresponding to the centroid of the perceptual loudness function |

The time-frequency features are referred to as either "global" or "local" features. The global features are computed using the Kliewer-Mertins (KM) defined start and end indices for all filter bank channels. In contrast, the local features define start and end indices for each sub-band by applying the KM technique to each sub-band. The features calculated from the pre-attack component are intended to quantify the spectral character of the pre-attack noise, which has been identified in the musical acoustics literature as being an important timbre-correlate. The pre-attack segment is defined by identifying the most significant attack using the KM algorithm – this is set as the end of the pre-attack segment – and then including the previous, pre-defined, number of samples [20] (in this case the pre-attack segment includes 128 samples). Thus, it is possible that the pre-attack noise segment overlaps either completely or partially with the vocalization defined by the KM technique in Section 2.1.1. The following table lists the names of all 58 one-dimensional features used by the aural classifier.

| Feature Number | Feature |
|---|---|
| 1 | Loudness centroid |
| 2 | Local minimum sub-band attack time |
| 3 | Frequency bin containing local minimum sub-band attack time |
| 4 | Local minimum sub-band attack slope |
| 5 | Frequency bin containing local minimum sub-band attack slope |
| 6 | Local mean sub-band attack time |
| 7 | Local mean sub-band attack slope |
| 8 | Local maximum sub-band attack time |
| 9 | Frequency bin containing local maximum sub-band attack time |
| 10 | Local maximum sub-band attack slope |

| Feature Number | Feature |
|---|---|
| 11 | Frequency bin containing local maximum sub-band attack slope |
| 12 | Global minimum sub-band attack time |
| 13 | Frequency bin containing global minimum sub-band attack time |
| 14 | Global minimum sub-band attack slope |
| 15 | Frequency bin containing global minimum sub-band attack slope |
| 16 | Global mean sub-band attack time |
| 17 | Global mean sub-band attack slope |
| 18 | Global maximum sub-band attack time |
| 19 | Frequency bin containing global maximum sub-band attack time |
| 20 | Global maximum sub-band attack slope |
| 21 | Frequency bin containing global maximum sub-band attack slope |
| 22 | Local minimum sub-band decay time |
| 23 | Frequency bin containing local minimum sub-band decay time |
| 24 | Local minimum sub-band decay slope |
| 25 | Frequency bin containing local minimum sub-band decay slope |
| 26 | Local mean sub-band decay time |
| 27 | Local mean sub-band decay slope |
| 28 | Local maximum sub-band decay time |
| 29 | Frequency bin containing local maximum sub-band decay time |
| 30 | Local maximum sub-band decay slope |
| 31 | Frequency bin containing local maximum sub-band decay slope |
| 32 | Global minimum sub-band decay time |
| 33 | Frequency bin containing global minimum sub-band decay time |
| 34 | Global minimum sub-band decay slope |
| 35 | Frequency bin containing global minimum sub-band decay slope |
| 36 | Global mean sub-band decay time |
| 37 | Global mean sub-band decay slope |
| 38 | Global maximum sub-band decay time |
| 39 | Frequency bin containing global maximum sub-band decay time |
| 40 | Global maximum sub-band decay slope |
| 41 | Frequency bin containing global maximum sub-band decay slope |
| 42 | Maximum sub-band correlation |
| 43 | Frequency bin containing maximum sub-band correlation |
| 44 | Mean sub-band correlation |
| 45 | Minimum sub-band correlation |
| 46 | Frequency bin containing minimum sub-band correlation |
| 47 | Psychoacoustic maxima-to-spectral-bins ratio |
| 48 | Psychoacoustic bin-to-bin difference |
| 49 | Duration |
| 50 | Peak loudness value |
| 51 | Peak loudness frequency |
| 52 | Pre-attack integrated loudness |
| 53 | Pre-attack peak loudness value |

| Feature Number | Feature |
|---|---|
| 54 | Pre-attack peak loudness frequency |
| 55 | Pre-attack psychoacoustic maxima to spectral bins ratio |
| 56 | Pre-attack psychoacoustic bin-to-bin difference |
| 57 | Pre-attack loudness centroid |
| 58 | Integrated loudness |