Title

Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California

Permalink
https://escholarship.org/uc/item/79q3p311

Journal
Journal of the Acoustical Society of America, 121(3)

Authors
Roch, Marie A
Soldevilla, Melissa S
Burtenshaw, Jessica C
et al.

Publication Date
2007-03-01

DOI
10.1121/1.2400663

Peer reviewed

Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California

Marie A. Roch[a)]
*Department of Computer Science, San Diego State University, 5500 Campanile Drive,
San Diego, California 92182-7720*

Melissa S. Soldevilla, Jessica C. Burtenshaw,
E. Elizabeth Henderson, and John A. Hildebrand
*Scripps Institution of Oceanography, The University of California at San Diego,
La Jolla, California 92093-0205*

(Received 12 May 2006; revised 16 October 2006; accepted 24 October 2006)

A method for the automatic classification of free-ranging delphinid vocalizations is presented. The vocalizations of short-beaked and long-beaked common (*Delphinus delphis* and *Delphinus capensis*), Pacific white-sided (*Lagenorhynchus obliquidens*), and bottlenose (*Tursiops truncatus*) dolphins were recorded in a pelagic environment of the Southern California Bight and the Gulf of California over a period of 4 years. Cepstral feature vectors are extracted from call data which contain simultaneous overlapping whistles, burst-pulses, and clicks from a single species. These features are grouped into multisecond segments. A portion of the data is used to train Gaussian mixture models of varying orders for each species. The remaining call data are used to test the performance of the models. Species are predicted based upon probabilistic measures of model similarity with test segment groups having durations between 1 and 25 s. For this data set, 256 mixture Gaussian mixture models and segments of at least 10 s of call data resulted in the best classification results. The classifier predicts the species of groups with 67%–75% accuracy depending upon the partitioning of the training and test data. *© 2007 Acoustical Society of America.* [DOI: 10.1121/1.2400663]

PACS number(s): 43.80.Ev, 43.60.Uv [WWA] Pages: 1737–1748

I. INTRODUCTION

Long-term acoustic monitoring is an established technique for assessing cetacean relative abundance and seasonality (Thompson and Friedl, 1982). Key steps in processing acoustic monitoring data are acoustic call detection and species classification. Identification of stereotyped mystecete calls has been accomplished using automatic detectors (e.g., Sirovic *et al.*, 2004), but odontocete call identification is more difficult owing to their calls' greater complexity. Species identification for odontocete calls has been accomplished using trained analysts, as well as automated classification based on similarity to calls collected in the presence of known species (Oswald *et al.*, 2003).

Recent advances in acoustic recording capabilities allow remote autonomous recordings with terabyte data storage (Wiggins, 2003). Manual analyses of these large data sets are prohibitively expensive. Reliable automated methods are needed for detection and classification of odontocete calls to allow rapid analysis of these large acoustic data sets.

Unlike many mammals (Fenton and Bell, 1981; Goold and Jones, 1995; Thompson *et al.*, 1992, 1996) and birds (Marler, 1957) which exhibit stereotyped calls that are readily distinguishable by species, delphinids have a wide and varied vocal repertoire that makes species identification more complex (Oswald *et al.*, 2003; Thompson and Richardson, 1995). Dolphin calls can be broken down into three general categories: echolocation clicks, burst-pulsed calls, and whistles (Popper, 1980). Each of these call types exhibits complex time- and frequency-varying features that differ across species and individuals. Echolocation clicks are broadband, impulsive sounds which typically range between 10 and 150 kHz in many dolphin species (Au, 1993). Echolocation clicks are used for prey-finding and navigation. Burst-pulsed calls are rapid series of broadband clicks which are not individually distinguishable to humans, resulting in calls with a scream-like, tonal quality (Murray *et al.*, 1988). These calls can range from 5 to 150 kHz and are thought to function for communicative purposes. Whistles are frequency modulated narrowband tonal calls which occur between 2 and 35 kHz (Thompson and Richardson, 1995). Whistles are thought to have communicative functions and it has been suggested that they may carry individual-specific information in some species (Caldwell *et al.*, 1990). While all dolphin species recorded to date produce click type calls, some species may not produce whistles (Herman and Tavolga, 1980).

Automatic classification of marine mammal calls involves at least three steps: signal detection, feature extraction, and classification. During signal detection, calls of interest are located within the larger time series. Feature extraction transforms each call to a feature vector or set of feature vectors which represents the salient characteristics of the call. Finally, the feature vectors are classified as belong-

[a)]Electronic mail: marie.roch@sdsu.edu ing to one of the target classes, or possibly as an unknown class. Target classes may be specific call types, individual animals, or species. The level of automation of each of these steps varies greatly in previously described cetacean classification studies.

Until recently, the standard approach has been to manually locate the end points of tonal calls and measure features such as fundamental frequency, harmonics, slope, and inflection points (e.g., Rendel *et al.*, 1999). More recent work, such as that of Oswald *et al.* (2005) can extract an expanded list of similar information automatically when given the start and end of a call. Datta and Sturtivant (2002) used edge detection techniques from computer vision to locate the whistle segments in a spectrogram.

Cepstral processing is a useful feature extraction technique used in human speech analysis. The real cepstrum is the discrete cosine transform (DCT) of the log of the short time spectral magnitude (Picone, 1993). When the source-filter model (Harrington and Cassidy, 1999) is assumed for the production of calls, this transformation results in the source information being typically contained in higher orders of the cepstrum. These can be discarded, resulting in a feature vector which captures information about the filter. While this will result in the loss of information about the source, it typically leads to reductions in both the amount of data needed to train effective models and in the computational time needed to train and use the classifier. Fitch (2000) notes that the source-filter model appears to be applicable for all mammals whose sound production has been studied. While odontocetes have different sound production systems from other mammals, the source-filter model is still a relevant paradigm (Cranford, 2000).

To further reduce the size of the feature vector, it is common to apply a set of filter banks that are spaced linearly at low frequencies and logarithmically at higher frequencies before computing the DCT. In bioacoustic studies of elephant calls and bird song (Clemins *et al.*, 2005; Kogan and Margoliash, 1998), a spacing based upon human psychophysics studies (the Mel scale, Sundberg, 1991) has been proposed. Due to differences in hearing, we believe that a more neutral approach or an approach specific to individual species such as the extensions to Hermansky's perceptual linear prediction (1990) proposed by Clemins and Johnson (2006) are appropriate. However, when working with multiple species, care must be taken to either design an aggregate filterbank or to perform the feature extraction for each species.

The classification of extracted feature vectors is accomplished using machine learning techniques. The majority of researchers have used algorithms that require supervised classification. These algorithms learn a partitioning of the feature space based upon training vectors. Once the classifier has been trained, feature vectors are assigned to classes based upon tests which determine to which partition the feature vectors belong.

One of the simplest supervised classifiers is linear discriminant analysis (Duda *et al.*, 2001), a technique which finds the hyperplane that best separates pairs of classes in a set of labeled training data. Steiner (1981) used linear discriminant analysis to differentiate the whistles of five dolphin species.

Other techniques use combinations of hyperplanes. Classification and regression trees (Duda *et al.*, 2001) is a related technique where multiple hyperplanes hierarchically partition the feature space into hypercubes. This technique has been used by Oswald *et al.* (2004) to determine the species of wild dolphin calls. Neural networks have been used by numerous groups. They are capable of separating the feature space into complex subregions associated with specific classes. A common method of integrating temporal data is to take a set of N evenly spaced samples from the feature data and to assemble them into a higher-dimensional feature vector. This approach has been used with backpropagation neural networks for the tasks of differentiating killer whale dialects, bottlenose dolphin clicks, and the detection of bowhead whale song notes (Deecke *et al.*, 1999; Houser *et al.*, 1999; Potter *et al.*, 1994). Alternative strategies to concatenating vectors are possible, such as the spectral averaging used by Potter *et al.* (1994).

When the goal is to recognize a specific call, the situation is complicated in that different repetitions of the call may be produced at different rates (Buck and Tyack, 1993). The rates of different portions of the call can vary considerably, and linear scaling is unlikely to capture the variation appropriately. One technique to cope with this is the use of dynamic time warping (DTW) (Rabiner and Juang, 1993), a dynamic programming technique that aligns the call to be classified to a reference call. This has been used to recognize whistles from a small set of captive dolphins, calls in captive bird song, and free ranging bowhead whales' calls (Buck and Tyack, 1993; Kogan and Margoliash, 1998; Mellinger and Clark, 2000). An alternate strategy is the use of hidden Markov models (HMMs) (Rabiner and Juang, 1993). HMMs are also capable of nonlinear time alignment, and have been shown both in the speech and bioacoustic communities to be effective classifiers. They have been used for distinguishing individual African elephants and their call types, dolphin group identities, bird individuals, and bowhead whale calls (Clemins *et al.*, 2005; Datta and Sturtivant, 2002; Kogan and Margoliash, 1998; Mellinger and Clark, 2000). In general, these provide a more robust performance than DTW, but HMMs require more data to estimate the model parameters.

A few research groups have studied unsupervised classifiers. Unsupervised classifiers attempt to learn classes from unlabeled data sets. Murray *et al.* (1988) used Kohonen's self-organizing maps and competitive learning to discern classes from a false killer whale call data set. With either technique, the goal is to have the network learn the similarities and differences between the feature vectors. Both methods were successful in learning a number of statistically homogeneous classes from calls produced by two individuals, and the authors were able to make links between the automatically discovered categories and those commonly given by humans such as clicks and whistles. Recently, Deecke and Janik (2006) combined adaptive resonance theory (ART) with DTW. In ART networks, a new pattern is compared to models for existing ones. If the new pattern differs suffi-

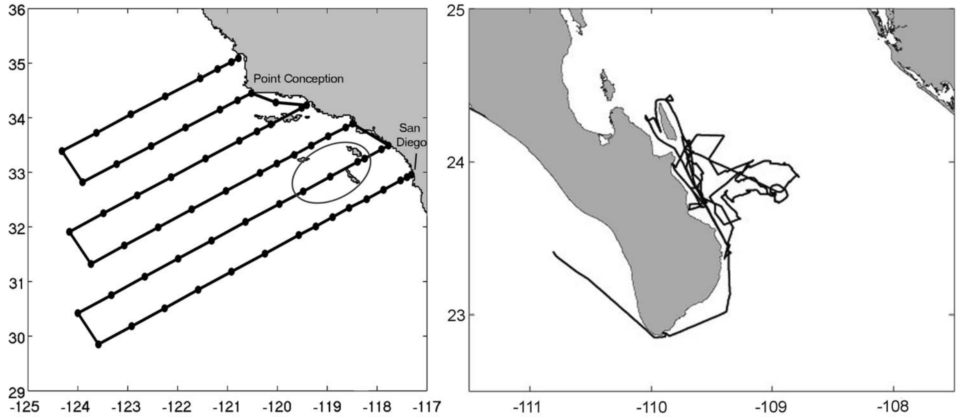

FIG. 1. Acoustic recordings were made along the transect and track lines in (a) the Southern California Bight and (b) the Gulf of California. The southern California data were collected on two series of cruises, one of which concentrated on the area contained in the square region.

ciently as determined by an experimenter controlled threshold, a new pattern is formed. DTW is used to provide the difference measure between existing patterns in the ART network and new ones.

Our study examines the performance of Gaussian mixture model (GMM) classifiers (Huang *et al.*, 2001) for determining the species of groups of free ranging dolphins recorded in the offshore Southern California Bight and the Gulf of California. GMMs are well known for their ability to model arbitrarily complex distributions with multiple modes and are effective classifiers for many tasks. They are functionally equivalent to ergodic hidden Markov models, and are appropriate for the species identification task where there is no expectation as to what component of a call will follow the current one and multiple calls may occur simultaneously. The classifier uses cepstral feature vectors and is able to distinguish the studied species accurately in 67%–75% of the test cases.

II. METHODS

The methods of this study are organized by task. We first describe the collection of the call data and its characteristics. The processing of the data is separated into call detection, feature extraction, and classification.

A. Data collection and species call descriptions

Acoustic recordings were collected on multiple cruises offshore of Southern California and within the Gulf of California between 2001 and 2005 (Fig. 1). Standard line-transect surveys were conducted to visually identify cetaceans in the study area. When a single-species school was encountered, an SSQ-57B sonobuoy was deployed, and the ship was positioned 1 to 2 km away from the school. SSQ-57B sonobuoys have a flat frequency response from 20 Hz to 20 kHz. The sonobuoy signal was transmitted to a multichannel receiver located on the ship. Acoustic data sampled at 48 kHz were recorded either directly to hard drive or to one of the following Sony DAT recorders: PCM-M1, TCD-D7, or TCD-D8. To improve the likelihood that recordings contain only single species call-types, we only analyzed recordings obtained when no other species' schools were sighted within 5 km of the sonobuoy location, and only included calls with high signal-to-noise ratio (SNR).

Four dolphin species that are commonly found and recorded in this region include short-beaked common (*Delphinus delphis*), long-beaked common (*Delphinus capensis*), Pacific white-sided (*Lagenorhynchus obliquidens*), and bottlenose (*Tursiops truncatus*) dolphins. As sighting logs for our early recordings did not distinguish the two species of common dolphins, the automatic classification system uses the *genus* for these animals. Bottlenose dolphins are the least abundant of the delphinids in our study area. When they are sighted, they are frequently in mixed groups with Risso's dolphins (*Grampus griseus*). Much of our bottlenose dolphin data are from the Gulf of California, where the same collection procedures were used with the exception of line-transect surveys. Other species known to inhabit the Southern California Bight, Risso's and northern right whale (*Lissodelphis borealis*) dolphins, were not encountered frequently enough to be included in the analysis.

Common dolphins produce whistles, burst pulses, and echolocation click trains (Au, 1993; Caldwell and Caldwell, 1968; Moore and Ridgway, 1995). Their whistles have a mean duration of 0.8 s, a mean minimum frequency of 7.4 kHz, a mean maximum frequency of 13.6 kHz, and a mean of 1.2 inflection points (Oswald *et al.*, 2003). Common dolphin clicks have source levels of 160–170 dB re 1 μPa at 1 m, pulse durations between 50 and 250 μs, and peak frequencies between 23 and 67 kHz (Au, 1993; Evans, 1973; Fish and Turl, 1975). Whistles made up the majority of calls we recorded (99%), with many of them overlapping, whereas burst pulses (1%) and the lower portion of their clicks were present at low numbers ($<1\%$). For all the species we recorded, percentages of click trains may be low as overlapping click trains were not distinguished.

Free-ranging bottlenose dolphins produce all three call types, with individual whistle characteristics including durations between 0.6 and 1.4 s, minimum frequencies between 5.4 and 8.5 kHz, maximum frequencies between 11.32 and 17.2 kHz, and 1.86 and 3.7 inflection points for a variety of populations (Acevedo-Gutiérrez and Stienessen, 2004; Oswald *et al.*, 2003; Steiner, 1981; Wang *et al.*, 1995). Bottlenose echolocation clicks have source levels of 228 dB re

TABLE I. Number of seconds of usable call data obtained for each dolphin species by date.

	Common		Pacific white-sided		Bottlenose	
Recording session	Date	s	Date	s	Date	s
1	30 April 2001	526	30 April 2001	398	15 April 2002	242
2	1 May 2001	159	20 June 2001	59	6 March 2004	350
3	4 November 2003	330	20 August 2003	401	8 March 2004	240
4					10 March 2004	409
5					15 May 2005	363
6					17 May 2005	488
7					18 May 2005	1133
8					19 May 2005	264
9					21 May 2005	349
Total		1015		858		3838

1 μPa at 1 m, pulse durations between 50 and 80 μs and peak frequencies between 110 and 130 kHz, though these may vary with location (Au, 1993). Our bottlenose dolphin recordings contained 81% whistles, 8% click trains, and 11% burst pulses.

Pacific white-sided dolphin echolocation clicks have been recorded with source levels of 170 dB re 1 μPa, pulse durations between 25 and 1000 μs, and peak frequencies between 50 and 80 kHz and 100 and 120 kHz (Evans, 1973; Fahner *et al.*, 2004; Nakamura and Akamatsu, 2004). While whistles have been recorded from Pacific white-sided dolphins (Caldwell and Caldwell, 1971; Whitten and Thomas, 2001), few were recorded during our sessions (21%) when compared to other species. The majority of Pacific white-sided calls we recorded were burst pulses (70%) and the lower frequency portion of click trains (10%).

B. Call detection

The detection of calls was accomplished manually. The start and end point of sets of whistles, burst pulses, and clicks were identified using spectrograms and audition when possible. Only calls which were deemed to be of sufficient quality as judged by SNR across the call bandwidth were used. Typically, these calls had SNRs of greater than 18 dB, and comprised approximately 65% of all detected calls. No effort was made to denote the start or end of individual calls, or to segregate individual calls from those that occurred with other conspecifics. Table I summarizes the call data used in this study by recording date.

Sessions 2–9 of the bottlenose call data were recorded in the Gulf of California; all other recordings were made in the Southern California Bight.

C. Feature extraction

Cepstral feature vectors were used to represent the short time spectrum of odontocete call data. No attempt was made to isolate individual calls, and the classifier learned the collection of sounds produced by groups of dolphins.

As shown in Fig. 2, the process consisted of computing the squared magnitude frequency response of a 21 ms frame which had been windowed with a Hamming window. A filter bank consisting of 64 linearly spaced overlapping triangular filters was applied between 5 and 23.5 kHz. The lower edge of 5 kHz was selected as the SNR for calls tended to be poor at frequencies beneath this threshold. The discrete cosine transform of the log filter bank outputs was computed, resulting in a 64 dimensional cepstral feature vector. These frames were computed every 11 ms, resulting in a 52% overlap be-

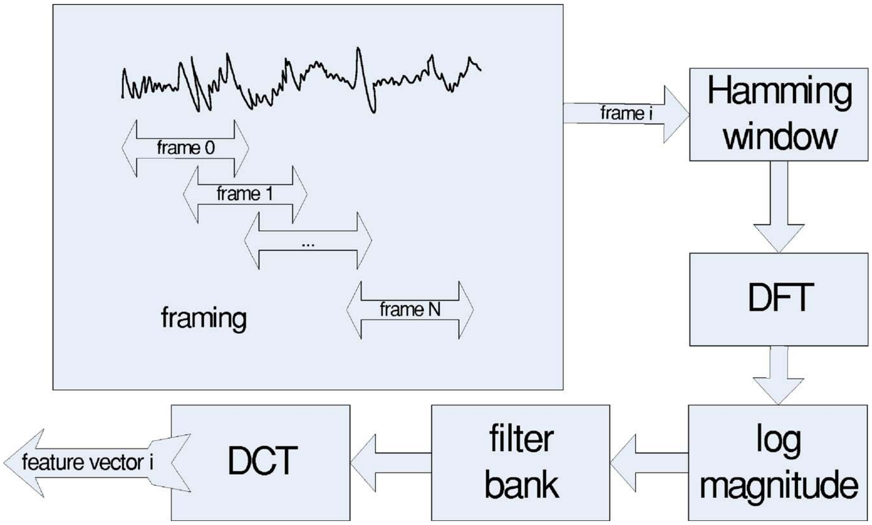

FIG. 2. Flow diagram for feature extraction. Overlapping frames of 21 ms are taken from the signal and transformed to the cepstral domain.

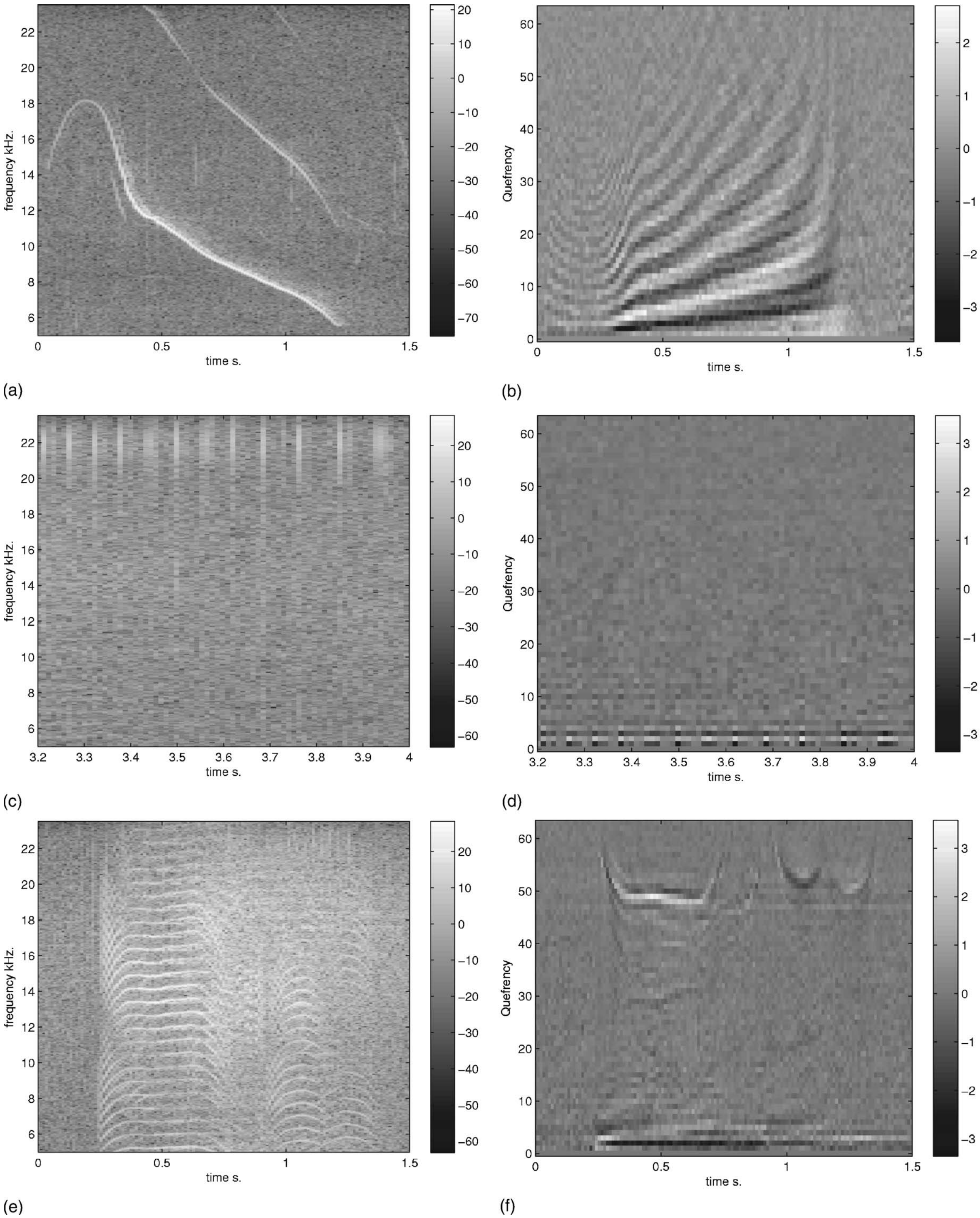

FIG. 3. Spectrograms and cepstral coefficients for (a) and (b) whistles, (c) and (d) clicks, and (e) and (f) burst pulses. The left column shows a spectrogram, and the right column shows the corresponding cepstrogram of the same signal after the application of a 64 point DFT based filter bank. The whistles are produced by bottlenose dolphins, the clicks and burst pulses by Pacific white-sided dolphins.

tween successive frames. Further details of this process can be found in Clemins *et al.* (2005) or Picone (1993).

Figure 3 shows spectrograms and their corresponding cepstrograms for each of the three call types. Cepstrograms are similar to spectrograms and display time series of cepstral vectors. In the whistle cepstrogram, one can see "harmonic" like structure in the cepstral domain related to the frequency modulated (FM) sweep. The harmonics move farther apart from one another as the frequency falls. Figures 3(c) and 3(d) show a spectrogram and cepstrogram for a click train. The location of the clicks is readily apparent in the cepstral domain and the majority of the information is concentrated in the lower quefrencies (cepstral coefficients). The burst pulse of Fig. 3(e) is also easily seen in Fig. 3(f).

Figure 3 shows information from single calls. In practice, many of the calls in the data set contain overlapping data. Figures 4(a) and 4(b) show the spectrogram and corresponding cepstrogram from a short segment of overlapping

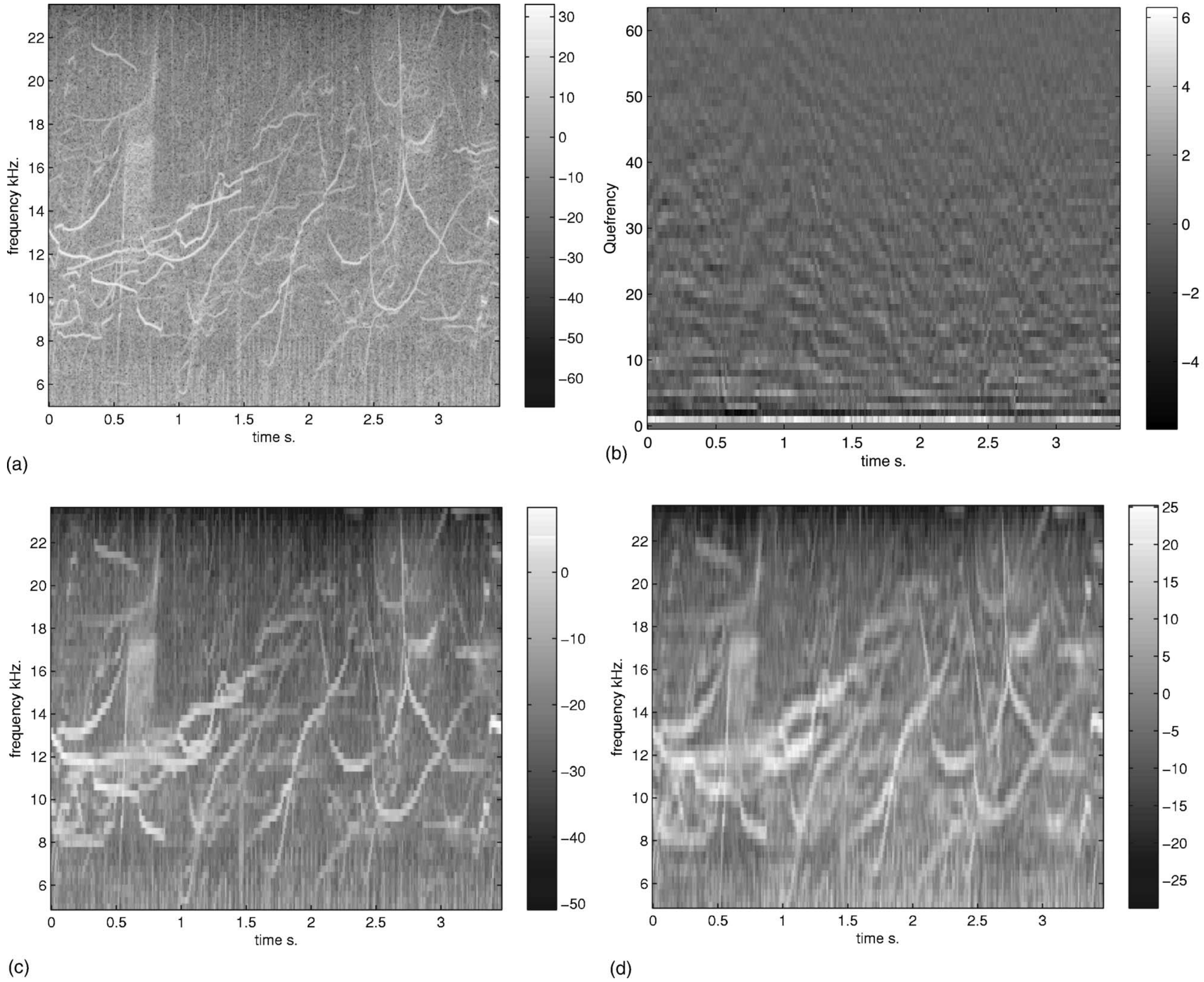

FIG. 4. Illustration of (a) spectrogram and (b) cepstrogram for multiple overlapping calls from a group of common dolphins. The classifier seeks patterns associated with multiple calls and does not attempt to separate out individual calls. The effects of a linearly spaced 64 band filterbank are shown in spectrogram (c). The cepstral series can be truncated while still retaining much of the original information as seen in the reconstruction of (c) using the first 32 cepstral coefficients of (b).

calls from common dolphins. To successfully classify this type of data, our strategy is to have the classifier learn the cepstral patterns of overlapping calls rather than attempting to isolate the individual calls.

The effect of the bandpass filtering operation on the same set of common dolphin calls can be seen in Fig. 4(c). As described in Sec. I, discarding part of the cepstrum can result in significant reductions of the feature space. This is advantageous as lower order models typically require less data to train. Higher quefrencies correspond to the fine detail in the log spectrum, and provide an opportunity to reduce the dimensionality of the feature space. To determine the effectiveness of the truncated cepstrum for representing delphinid calls, we reconstructed spectrograms by inverting the operations used to form the cepstrum. It was determined that retaining the first 32 frequencies resulted in spectrograms where major features of the call were still clearly evident as illustrated in Fig. 4(d).

An additional step of the feature extraction is to apply cepstral means subtraction which detrends the cepstrum by subtracting the mean vector. This operation has the dual effects of removing the mean from the log spectrum (Hermansky, 1995) and removing any constant contribution to the cepstrum caused by the convolution of the signal and the hydrophone. The removal of the hydrophone-specific contribution is critical when there are mismatches between the frequency responses of hydrophones used in the training and test sets. As all hydrophones used in this study were from SSQ-57B sonobuoys, it is assumed that the primary benefit of using cepstral means subtraction is to detrend the log spectrum.

The cepstral coefficients yield a static representation of the short term spectrum. Many audio classification tasks benefit from adding information about how the spectrum is changing, and this can be done by taking the first and second derivatives of the cepstrum. These were appended to the feature vector and improved the accuracy by approximately 20%.

D. Classification

The feature vectors were classified using GMMs. A GMM M consists of N normal distributions with mean μ_i and covariance matrix Σ_i where $1 \leqslant i \leqslant N$. Each of these distributions is scaled by c_i $(1 \leqslant i \leqslant N)$ and the sum of the c_i's must be one to ensure that the GMM represents a probability distribution.

The likelihood of each d dimensional observation x can be found by the following (Huang *et al.*, 2001):

$$\Pr(x|M) = \sum_{i=1}^{N} c_i \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \times \exp\left(-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right), \qquad (1)$$

where the prime $(')$ denotes the transpose operator and $|\cdot|$ the determinant.

The parameters for GMMs cannot be estimated using closed form equations. As the contribution of each training vector to specific mixtures is unknown, standard estimation techniques such as maximum likelihood estimation are not possible. However, if one assumes the existence of a model which approximates the distribution, it is possible to use an iterative algorithm and the training data to find a new model which better approximates the training data. One way of forming the initial model (Young *et al.*, 2002) is to start with a Gaussian classifier (equivalent to a single mixture GMM with $c_1 = 1$) that can be estimated by using the sample mean and covariance. The single mixture is split into two mixtures with identical covariances and means offset by 0.2 s.d. Each of the new mixtures is assigned half of the old mixture's weight. The iterative algorithm is typically executed for a few iterations and then the mixtures are split again. This is repeated until the desired number of mixtures is formed.

The iterative algorithm is an application of the expectation maximization (EM) algorithm (Moon, 1996). The central idea is that while it is not known which of the N Gaussian distributions is responsible for generating the t^{th} observation x_t, we can use the expected value as an estimator. This is represented by the notation γ^t_m, which is the expected contribution of the m^{th} mixture to the total likelihood associated with observation x_t in the i^{th} version of the model:

$$\gamma^t_m = \frac{c_m \Pr(x_t|\mu_m^{(i)}, \Sigma_m^{(i)})}{\Pr(x_t|M^{(i)})}. \qquad (2)$$

With the expectation known, maximum likelihood techniques can be used to produce the next iteration of the model $M^{(i+1)}$:

$$c_m^{(i+1)} = \frac{\Sigma_{t=1}^{T} \tau^t_m}{N}, \qquad (3)$$

$$\mu_k^{(i+1)} = \frac{\Sigma_{t=1}^{T} \tau^t_m x_t}{\Sigma_{t=1}^{T} \tau^t_m}, \qquad (4)$$

$$\sum\nolimits_k^{(i+1)} = \frac{\Sigma_{t=1}^{T} \tau^t_m (x_t - \mu^{(i)})(x_t - \mu^{(i)})'}{\Sigma_{t=1}^{T} \tau^t_m}. \qquad (5)$$

At each iteration, the likelihood of the training data with respect to the new model is guaranteed to be greater than or equal to the likelihood with respect to the old model. Thus the EM algorithm will converge to a local maximum. While there are no known proofs of the rate of convergence, convergence is typically fast with anywhere from 5 to 15 iterations. The derivation of these equations can be found in Huang *et al.* (2001).

A number of assumptions were made with respect to our use of GMMs. Iteration was stopped when the likelihood of the new model was no more than 2% greater than the previous one. Based upon the asymptotic independence of the components of cepstral feature vectors (Merhav and Lee, 1993), it was also assumed that the components of the feature vectors were independent. This resulted in diagonal covariance matrices which significantly reduced the computational cost.

Once the models were trained, the posterior probability of each species was computed with respect to a set of test vectors:

$$\Pr(\text{species}|\text{test}) = \frac{\Pr(\text{test}|\text{species})\Pr(\text{species})}{\Pr(\text{test})}. \qquad (6)$$

The right-hand side of Eq. (6) was obtained from Bayes rule. The class label for the test segment was decided using Bayes decision rule (Duda *et al.*, 2001), which selected the class that produces the maximum probability.

In this work, a uniform prior distribution was assumed, resulting in Pr(species) having a constant contribution to each posterior probability. As Pr(test) was also constant across species, the maximum posterior probability relied solely on the class conditional likelihood, Pr(test|species), which was evaluated by Eq. (1). As observations were assumed to be independent from one another, the log likelihoods from each observation were summed to produce the joint posterior likelihood as shown in Fig. 5.

The GMMs were implemented using the Hidden Markov Model Toolkit (HTK) by Young *et al.* (2002), an open source suite of programs for speech recognition. Customizations were made to support the linear filter bank. A series of control programs were written in PYTHON, a general purpose object oriented scripting language.

Experiments were conducted to examine the effect of model order, length of training and test data, and variations of the choice of training data. Table II indicates the partitioning of call data from Table I into training and test data. With the exception of the experiments which examined the choice of training data, all experiments used partition 1. Two sessions were selected for the bottlenose data simply due to the abundance of data available. The common dolphin data from session 1 were not used as training data as they represented the longest session. Using this session would have reduced the number of test cases resulting in an increase of the 95% confidence interval (CI).

When designing the partitions, the authors attempted to minimize the risk that calls from the same animals in the same behavior state were contained in both the training and test sets. On some of the cruises, ship track patterns may have allowed resampling the same group of dolphins on the same day, so sessions from a single day are never split into training and test data. Given the fission-fusion nature of dolphin school groupings (Connor *et al.*, 2000; Neumann, 2001)

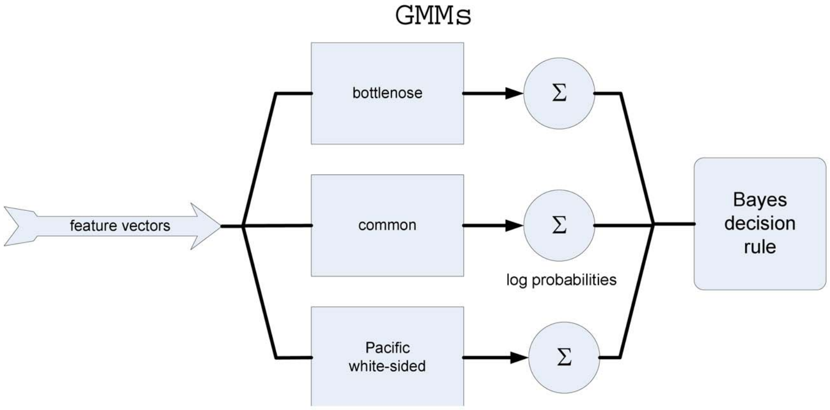

FIG. 5. Classification of sequences of feature vectors by Gaussian mixture models.

along with changes in behavior state, the authors believe that using recordings from different days is a conservative approach to reducing the likelihood of encountering the same group of animals in the same behavior state. Due to the limited number of recording days, a separate evaluation set was not established.

The first set of experiments tested model order. GMMs with 64, 128, 256, and 512 mixtures were trained. The mixture splitting algorithm made powers of two a logical choice for the model order. Test data were partitioned into 20 s segments that were classified as a single unit regardless of the number of calls contained therein.

Test segment length was examined by using 1, 3, 5, 10, 15, 20, 25, and 30 s segments for the best model order associated with this data set. The effect of varying the quantity of training data was also explored. The common dolphin data from session 2 was the shortest data set with 159 s of training data, and all training length tests were designed to be shorter, with tests at 30, 60, 90, and 120 s of training data. The remaining training data were not added to the test set as this would have violated the constraint that calls from the same day should not serve as both training and test data. The bottlenose dolphin data were split between sessions 1 and 5 evenly, and tests were performed with 20 s segments of call data.

To ensure that the results were not overly dependent upon the specific calls used in the training set, three additional experiments were performed with 256 mixture GMMs and 10 s test segments. In each experiment, the training data for one of the groups was substituted with a different set of training data. When feasible, training data of a similar length as in the original set of experiments were used. The test sets were appropriately updated by deleting the new training data and inserting the old. Table II provides a listing of the sessions used to produce these three new partitions (2–4) of the data set.

TABLE II. Selection of train/test data. The listed sessions from Table I are used for training with the remaining sessions used as test. Partitions 2–4 were chosen such that the training data for a single species is replaced by one of the other sessions.

Partition	Common	Pacific white-sided	Bottlenose
1	2	1	1, 5
2	3	1	1, 5
3	2	3	1, 5
4	2	1	8, 9

Overall accuracy was defined as the percentage of test segments that were correctly identified. If the correctness of the outcome of each test segment is considered as a binomial trial, it is possible to construct a 95% CI for the mean (Huang *et al.*, 2001). The confidence interval is defined as follows:

$$\mathrm{CI}(\alpha,p,N) = \pm F_n^{-1}\left(1 - \frac{\alpha}{2}\,\middle|\, 0, \sqrt{\frac{p(1-p)}{N}}\right), \qquad (7)$$

where $F_n^{-1}(\cdot|\mu,\sigma)$ denotes the inverse cumulative distribution function of a normal distribution with mean μ and standard deviation σ. The confidence interval is controlled by α, which is set to one minus the desired confidence interval (0.05 for the 95% CI). The variable p denotes the accuracy, and N is the number of trials.

A second statistic was defined to prevent the large quantity of bottlenose dolphin calls from biasing the results toward classifiers that favor that species. The average of the per species accuracies was calculated where per species accuracy was defined as the number of correct classifications of each species divided by the number of species-specific classification attempts. Large deviations in this statistic from the overall accuracy are indicative of a classifier bias toward a specific species.

III. RESULTS

The accuracy with respect to the number of mixtures is reported in Fig. 6 where the number of mixtures varies between 64 and 512 by powers of 2. The circles represent the percentage of correctly classified segments from all species, and the error bars show the 95% CI for each test. The classifier accuracy increases as the number of mixtures climbs, with a maximum accuracy of 78.1% with 512 mixtures. The average per species accuracies are plotted with triangles, and reported by species in Fig. 7. With respect to species-specific performance, varying the number of mixtures resulted in trade-offs between common and bottlenose dolphin performance. Pacific white-sided dolphins were always well classified.

The 256 mixture models provided the best balance of overall and per species accuracy, and all subsequent experi-

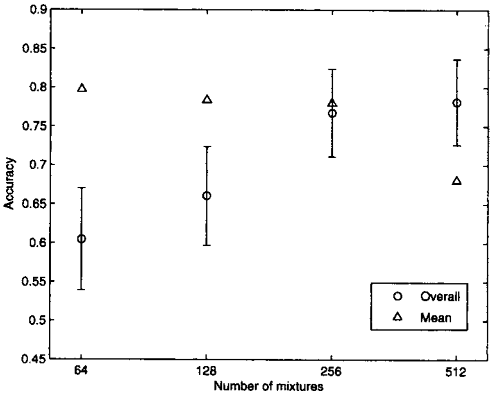

FIG. 6. Classification accuracy using Gaussian mixture models with differing numbers of mixtures. The circles represent the percentage of segments correctly classified and their 95% confidence intervals. The triangles show the means of the classification rates across species. Considering only overall accuracy would bias the classifier toward the species with the greatest number of test utterances.

ments were conducted with 256 mixture GMMs. The overall and species-specific results of varying the test length between 1 and 30 s are summarized in Figs. 8 and 9. Accuracy increased as a function of test segment length up to 10 s. When test segments were longer than 10 s, any further increases generally fell within the 95% confidence interval for means.

The effect of training data length was also investigated with test segments of 20 s. As shown in Table III, reducing the amount of training data impacted both the overall and mean species accuracies. It was of note that the recognition rate for common dolphins was actually higher with shorter amounts of training data and this is discussed in the next section.

The final set of experiments examined the effect of varying the training and test partitions and is reported in Table IV. The experiments on partitions 2 and 3 had overall accuracies that were within the 95% confidence interval of partition 1. The partition 3 experiment showed a marked decrease in the accuracy of identifying Pacific white-sided dolphins, with 10 of the 41 tests incorrectly identified as common dolphins. The partition 4 experiment had a lower accuracy (67.1%).

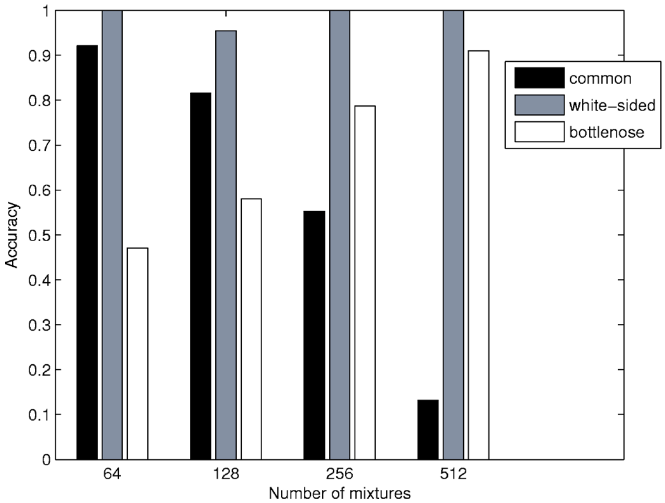

FIG. 7. Accuracy by species as the number of mixtures per GMM varies.

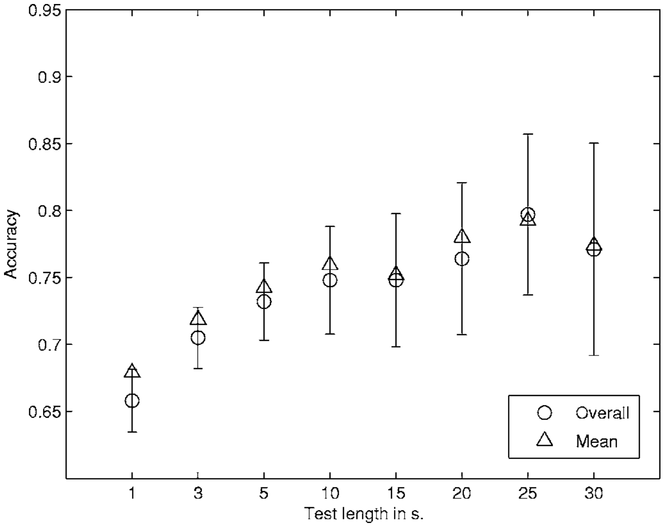

FIG. 8. Classification accuracy when varying the length of test segments classified by 256 mixture GMMs. The circles represent the percentage of segments correctly classified and their 95% confidence intervals. The triangles show the means of the classification rates across species.

IV. DISCUSSION

The experiments which varied the number of mixtures (Fig. 6) suggest that the optimal number of mixtures may vary from one species to another. Pacific white-sided dolphins were always recognized well. This may be related to their calls having less in common with the other species, including a substantially lower number of whistles. The primary trade off appeared to be between common and bottlenose dolphins with bottlenose dolphins better recognized by high order models and common dolphins by low order models. This may be due to the greater complexity of the bottlenose call repertoire, but could also be attributed to differences in the quantity of training data.

When the length of the test segment was varied (Figs. 8 and 9), increases in accuracy were correlated with the test length. However, with the exception of the experiment with a 25 s test length (accuracy: overall 79.7%±6.0%, species mean 79.2%), overall accuracy was contained within the 95% confidence interval of the 10 s test segments. Test segments longer than 10 s did not generally contribute significantly to increased accuracy. As the length of each test segment was increased, the total number of tests decreased. This accounts for the increased spread of the 95% confidence intervals seen in Fig. 8. Unlike the mixture experiment, the per species accuracies generally followed the same trend as the overall accuracy.

Varying the quantity of training data showed that accuracies of above 70% could be seen with as little as 90 s of training data per species. However, this should be taken with a caveat as an equally important question is related to the number of behavior states and environments from which the researcher collects data. One anomalous result within this set was the reduction of accuracy in the common dolphin data with 60 s of training data. This result is difficult to interpret, but may represent similarities between behavior states in different species. If the behavior state occurs in the training data of one of the other species but not in the common dolphin training data, it may classify to the other species. Another possible explanation is that with reduced training data, there is the danger of learning the environment as opposed to the species. Cepstral means subtraction will reduce the possibility of this occurring for stationary events, but not for other characteristics of the auditory scene.

Our examination of the training data variation experiments revealed that the Pacific white-sided dolphin data in session 3 contained brief but periodic signal drop outs due to ship board radio interference with the sonobuoy signals. This may contribute to the decreased accuracy when these data are used as training material (during test, the drop outs are small portions of any 10 s segment and are less likely to play a large role in the overall likelihood). Alternatively, it may be that the Pacific white-sided call set in partition 1 has more variation in the call data than that of partition 3.

The experiment with partition 4 was the only experiment that did not fall within the 95% confidence interval of the first partition's overall accuracy. The lower accuracy of 67.1% ±4.4% was due entirely to bottlenose dolphins which were misclassified as common dolphins. As with the decrease in Pacific white-sided dolphins, it is possible that the training set was less diverse. In addition, this experiment is the only experiment that used bottlenose training data exclusively from the Gulf of California and tested using data from both the Gulf of California and the Southern California Bight. An examination of the error rate of calls recorded in the Southern California Bight showed a below average accuracy of 33.3%, but the number of test segments ($N=24$) from a single day's recordings is too small to draw any conclusions about possible dialectal differences between the two groups. It should also be noted that other segments were misclassified. The 16 misclassifications of the Southern California Bight population were in a larger context of 105 bottlenose misclassifications. Recalling that the bottlenose data set is much larger than that of the other species, it should be noted that the mean accuracy of the three species is similar to the other partition tests.

Fluctuations in accuracy are to be expected with different partitions of the call data. Nonetheless, classification accuracy for three of the four tests fell within each other's 95% confidence interval for means. In all cases, classification accuracy was well above chance levels of guessing using a uniform prior (33.3%).

Finally, it should be noted that some of the choices made for feature selection were the result of compromises and will be the subject of future work. The current analysis window of 21 ms is too long for clicks which can be as short as 40 μs in bottlenose dolphins (Au, 1993). Similarly, the truncation of the cepstral series to 32 coefficients results in the loss of the inclick interval for burst pulses, which may be of relevance in species identification. As inclusion of the interclick intervals would have resulted in a significant increase of the dimensionality of the feature vector, a better strategy to include this may be to estimate the interclick interval separately and include it as one additional component of the feature vector. These issues are the subject of future work, but the current system still proves to be an effective classifier.

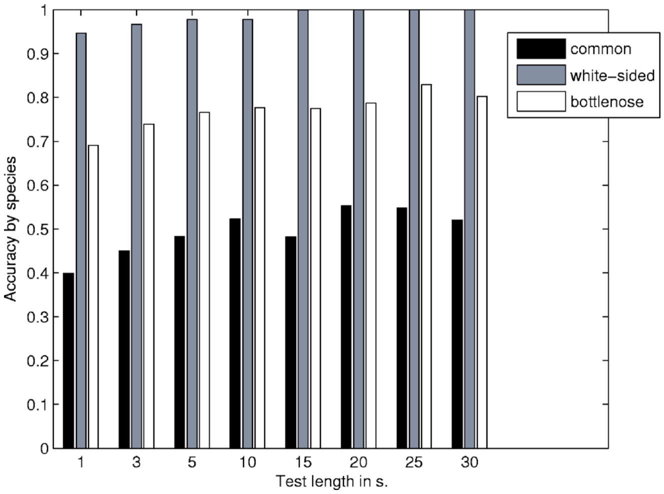

FIG. 9. Accuracy by species as the length of the test segment classified by a 256 mixture GMM is varied.

V. CONCLUSION

A system has been presented for the classification of free ranging dolphin calls that functions with an accuracy which generally ranges from 67% to 75% with a 95% CI of approximately ±4%. The system does not rely on specific call types and there is no requirement to separate out individual calls from the aggregate, which can be difficult with large groups of highly social and vocal animals. Using a cepstral feature space allows the system to capture the timbre of the calls, which is lost in systems that extract simple parameters from whistle curves. The use of supervised statistical learning permits the system to be used for other species and tasks with little modification.

We have not discussed comparisons between our system's performance and that of other call recognition systems. In addition to differences between the detection, feature extraction, and classification approach that each method uses, there are a number of parameters that make direct comparisons between methods tenuous at best. The vast differences

TABLE III. Performance of 256 mixture GMMs with 20 s test segments and varying amounts of training data per species.

Train s	Overall accuracy	±95% CI	Species mean	Common	Pacific white-sided	Bottlenose
30	0.512	0.067	0.682	0.842	0.818	0.387
60	0.735	0.059	0.666	0.158	1.000	0.839
90	0.767	0.057	0.714	0.447	0.864	0.832
120	0.725	0.060	0.741	0.474	1.000	0.748

TABLE IV. Results of 256 mixture GMM tests with 10 s test segments on the partitions resulting from Table II. Column CI is the 95% confidence interval on the overall accuracy, and species mean represents the mean of the individual species' accuracies.

Partition	Test count	Overall accuracy	±95% CI	Species mean	Common	Pacific white-sided	Bottlenose
1	449	0.748	0.040	0.759	0.523	0.978	0.777
2	433	0.707	0.043	0.739	0.557	0.956	0.704
3	447	0.729	0.041	0.690	0.523	0.767	0.780
4	447	0.671	0.044	0.723	0.523	0.978	0.668

in collection methods and difficulty of specific corpora have led human speech classification researchers to offer yearly competitions with common tasks and data sets (e.g. Przybocki and Martin, 2001).

While the great variety of bioacoustic classification tasks make the establishment of standards difficult, the authors believe that members of the bioacoustics community should continue to take steps to establish common data (Gaunt *et al.*, 2005) and software repositories which will permit direct comparison of classification algorithms.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their comments and suggestions as well as Ted Cranford for his helpful discussions on dolphin vocalizations. This work was supported by US Navy Chief of Naval Operations N45 – Frank Stone and Ernie Young, the Strategic Environmental Research and Development Project–Robert Holst, the Office of Naval Research–Bob Gisiner and Ellen Livingston, and US Government Award No. 0016.

Acevedo-Gutiérrez, A., and Stienessen, S. C. (**2004**). "Bottlenose dolphins (*Tursiops truncatus*) increase number of whistles when feeding," Aquat. Mamm. **30**, 357–362.

Au, W. W. L. (**1993**). *The sonar of Dolphins* (Springer, New York).

Buck, J. R., and Tyack, P. L. (**1993**). "A quantitative measure of similarity for *Tursiops truncatus* signature whistles," J. Acoust. Soc. Am. **94**, 2497–2506.

Caldwell, M. C., and Caldwell, D. K. (**1968**). "Vocalization of naive captive dolphins in small groups," Science, **159**, 1121–1123.

Caldwell, M. C., and Caldwell, D. K. (**1971**). "Stastistical evidence for individual signature whistles in Pacific whitesided dolphins, *Lagenorhynchus obliquidens*," Cetology **3**, 1–9.

Caldwell, M. C., Caldwell, D. K., and Tyack, P. L. (**1990**). "Review of the signature-whistle hypothesis for the Atlantic bottlenose dolphin," in *The Bottlenose Dolphin*, edited by S. Leatherwood and R. R. Reeves (Academic, San Diego).

Clemins, P. J., and Johnson, M. T. (**2006**). "Generalized perceptual linear prediction feature for animal vocalization analysis," J. Acoust. Soc. Am. **120**, 527–534.

Clemins, P. J., Johnson, M. T., Leong, K. M., and Savage, A. (**2005**). "Automatic classification and speaker identification of African elephant (*Loxodonta africana*) vocalizations," J. Acoust. Soc. Am. **117**, 956–963.

Connor, R. C., Wells, R. S., Mann, J., Read, A. J., Tyack, P. L., and Whitehead, H. (**2000**). "The bottlenose dolphin: Social relationships in a fission-fusion society," in *Cetacean Societies: Field Studies of Dolphins and Whales*, edited by J. Mann *et al.* (University of Chicago Press, Chicago), pp. 91–126.

Cranford, T. W. (**2000**). "In search of impluse sound sources in odontocetes," in *Hearing by Whales and Dolphins*, edited by W. W. L. Au, A. N. Popper, and R. R. Fay (Springer, New york), pp. 109–155; private communication (**2005**).

Datta, S., and Sturtivant, C. (**2002**). "Dolphin whistle classification for determining group identities," Signal Process. **82**, 127–327.

Deecke, V. B., Ford, J. K. B., and Spong, P. (**1999**). "Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale (*Orcinus orca*) dialects," J. Acoust. Soc. Am., **105**, 2499–2507, 2006.

Deecke, V. B., and Janik, V. M. (**2006**). "Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls," J. Acoust. Soc. Am. **199**, 645–653.

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

Evans, W. E. (**1973**). "Echolocation by marine delphinids and one species of freshwater dolphin," J. Acoust. Soc. Am. **54**, 191–199.

Fahner, M., Thomas, J., Ramirez, K., and Boehm, J. (**2004**). "Echolocation in bats and dolphins," in *Acoustic Properties of Echolocation Signals by Captive Pacific White-Sided Dolphins (Lagenorhynchus obliquidens)*, edited by J. Thomas, C. Moss, and M. Vater (University of Chicago Press, Chicago).

Fenton, M. B., and Bell, G. P. (**1981**). "Recognition of species of insectivorous bats by their echolocation calls," J. Mammal. **62**, 233–243.

Fish, J. E., and Turl, C. W. (**1975**). "Source level of four species of small toothed whales," in Conference on the Biology and Conservation of Marine Mammals, Santa Cruz.

Fitch, W. T. (**2000**). "The evolution of speech: A comparative review," Trends Cogn. Sci. **4**, 258–267.

Gaunt, S. L. L., Nelson, D. A., Dantzker, M. S., Budney, G. F., and Bradbury, J. W. (**2005**). "New directions for bioacoustics collections," Auk **122**, 984–988.

Goold, J. C., and Jones, S. E. (**1995**). "Time and frequency-domain characteristics of sperm whale clicks," J. Acoust. Soc. Am. **98**, 1279–1291.

Harrington, J., and Cassidy, S. (**1999**). *Techniques in Speech Acoustics* (Kluwer Academic, Dordrecht).

Herman, L. M., and Tavolga, W. N. (**1980**). "The communication system of cetaceans," in *Cetacean Behavior: Mechanisms and Functions*, edited by L. M. Herman (Wiley Interscience, New York), pp. 149–209.

Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. **87**, 1738–1752.

Hermansky, H. (**1995**). "Exploring temporal domain for robustness in speech recognition," in *The 15th International Congress on Acoustics*, Trondheim, Norway, Vol. **2**, pp. 61–64.

Houser, D. S., Helweg, D. A., and Moore, P. W. (**1999**). "Classification of dolphin echolocation clicks by energy and frequency distributions," J. Acoust. Soc. Am. **106**, 1579–1585.

Huang, X., Acero, A., and Hon, H.-W. (**2001**). *Spoken Language Processing* (Prentice Hall PTR, Upper Saddle River, NJ).

Kogan, J. A., and Margoliash, D. (**1998**). "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," J. Acoust. Soc. Am. **103**, 2185–2196.

Marler, P. (**1957**). "Specific distinctiveness in the communication signals of birds," Behaviour **11**, 13–39.

Mellinger, D. K., and Clark, C. W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation," J. Acoust. Soc. Am. **107**, 3518–3529.

Merhav, N., and Lee, C.-H. (**1993**). "On the asymptotic statistical behavior of empirical cepstral coefficients," IEEE Trans. Signal Process. **41**, 1990–1993.

Moon, T. K. (**1996**). "The expectation-maximization algorithm," IEEE Signal Process. Mag. **13**, 47–60.

Moore, S. E., and Ridgway, S. H. (**1995**). "Whistles produced by common dolphins from the Southern California Bight," Aquat. Mamm. **21**, 55–63.

Murray, S. O., Mercado, E., and Roitblat, H. L. (**1988**). "The neural network classification of false killer whale (*Pseudorca crassidents*) vocalizations," J. Acoust. Soc. Am. **104**, 3626–3633.

Nakamura, K., and Akamatsu, T. (**2004**). "Comparison of click characteristics among odontocete species," in *Echolocation in Bats and Dolphins*, edited by J. Thomas, C. Moss, and M. Vater (University of Chicago Press, Chicago).

Neumann, D. R. (**2001**). "The activity budget of free-ranging common dolphins (*Delphinus delphis*) in the northwestern Bay of Plenty, New Zealand," Aquat. Mamm. **27**, 121–136.

Oswald, J., Rankin, S., and Barlow, J. (**2004**). "The effect of recording and analysis bandwidth on acoustic identification of delphinid speicies," J. Acoust. Soc. Am. **116**, 3178–3185.

Oswald, J. N., Barlow, J., and Norris, T. F. (**2003**). "Acoustic identification of nine delphinid species in the eastern tropical Pacific Ocean," Marine Mammal Sci. **19**, 20–37.

Oswald, J. N., Rankin, S., Barlow, J., and Lammers, M. O. (**2005**). "A new tool for real-time acoustic species identification of delphinid whistles," J. Acoust. Soc. Am. **118**, 1909 (abstract only).

Picone, J. W. (**1993**). "Signal modeling techniques in speech recognition," Proc. IEEE **81**, 1215–1247.

Popper, A. N. (**1980**). "Sound emission and detection by delphinids," in *Cetacean Behavior: Mechanisms and Functions*, edited by L. M. Herman (Krieger, Malabar, FL).

Potter, J. R., Mellinger, D. K., and Clark, C. W. (**1994**). "Marine mammal call discrimination using artificial neural networks," J. Acoust. Soc. Am. **96**, 1255–1262.

Przybocki, M. A., and Martin, A. F. (**2001**). "The NIST speaker recognition evaluation: 1996-2001," in *Proceedings of the Odyssey*, Anogia, Crete, Greece.

Rabiner, L. R., and Juang, B.-H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).

Rendell, L. E., Matthews, J. N., Gill, A., Gordon, J. C. D., and Macdonald, D. W. (**1999**). "Quantitative analysis of tonal calls from five odontocete species, examining interspecific and intraspecific variation," J. Zool. 249, 403–410.

Sirovic, A., Hildebrand, J. A., Wiggins, S. M., McDonald, M. A., Moore, S. E., and Thiele, D. (**2004**). "Seasonality of blue and fin whale calls and the influence of sea lee in the Western Antarctic Peninsula," Deep-Sea Res., Part II, **51**, 2327–2344.

Steiner, W. W. (**1981**). "Species-specific differences in pure tonal whistle vocalizations of five western north Atlantic dolphin species," Behav. Ecol. Sociobiol. **9**, 241–246.

Sundberg, J. (**1991**). *The Science of Musical Sound* (Academic, San Diego).

Thompson, D. H., and Richardson, W. J. (**1995**). "Marine mammal sounds," in *Marine Mammals and Noise*, edited by W. J. Richardson, C. R. Greene, Jr., C. I. Malme, and D. H. Thomson (Academic, San Diego), pp. 325–386.

Thompson, P., Findley, L. T., and Vidal, O. (**1992**). "20 Hz pulses and other vocalizations of fin whales, *Balaenoptera physalus*, in the Gulf of California, Mexico," J. Acoust. Soc. Am. **92**, 3051–3057.

Thompson, P. O., Findley, L. T., Vidal, O., and Cummings, W. C. (**1996**). "Underwater sounds of blue whales, *Balaenoptera musculus*, in the Gulf of California, Mexico," Marine Mammal Sci. **12**, 288–293.

Thompson, P. O., and Friedl, W. A. (**1982**). "A long term study of low frequency sound from several species of whales off Oahu, Hawaii," Cetology **45**, 1–19.

Wang, D., Wursig, B., and Evans, W. (**1995**). "Comparisons of whistles among seven odontocete species," in *Sensory Systems of Aquatic Mammals*, edited by J. A. Kastelein, R. A. Thomas, and P. E. Nachtigall (De Spil, Woerden, NL).

Whitten, J. L., and Thomas, J. A. (**2001**). "Whistle repertoire of Pacific white-sided dolphins (*Lagenorhynchus obliquidens*) at the John G. Shedd Aquarium," J. Acoust. Soc. Am. **109**, 2391 (abstract only).

Wiggins, S. (**2003**). "Auonomous acoustic recording packages (ARP's) for long-term monitoring of whale sounds," Mar. Technol. Soc. J. **37**, 13–22.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (**2002**). The HTK book, version 3.2, URL http://htk.eng.cam.ac.uk. Last viewed 2/8/07.