Using Gaussian mixture models to detect and classify dolphin whistles and pulses

Pablo Peso Parada[a] and Antonio Cardenal-López
Multimedia Technologies Group, Department of Signal Processing and Communications, University of Vigo, 36310 Vigo, Spain

(Received 8 March 2013; revised 22 April 2014; accepted 2 May 2014)

In recent years, a number of automatic detection systems for free-ranging cetaceans have been proposed that aim to detect not just surfaced, but also submerged, individuals. These systems are typically based on pattern-recognition techniques applied to underwater acoustic recordings. Using a Gaussian mixture model, a classification system was developed that detects sounds in recordings and classifies them as one of four types: background noise, whistles, pulses, and combined whistles and pulses. The classifier was tested using a database of underwater recordings made off the Spanish coast during 2011. Using cepstral-coefficient-based parameterization, a sound detection rate of 87.5% was achieved for a 23.6% classification error rate. To improve these results, two parameters computed using the multiple signal classification algorithm and an unpredictability measure were included in the classifier. These parameters, which helped to classify the segments containing whistles, increased the detection rate to 90.3% and reduced the classification error rate to 18.1%. Finally, the potential of the multiple signal classification algorithm and unpredictability measure for estimating whistle contours and classifying cetacean species was also explored, with promising results.
 [http://dx.doi.org/10.1121/1.4876439]

PACS number(s): 43.30.Sf, 43.60.Bf, 43.60.Lq [WWA] Pages: 3371–3380

I. INTRODUCTION

Recent years have seen a growing interest in the study of cetaceans using underwater recordings. Cetaceans belonging to the mysticete (baleen whales) and odontocete (toothed whales, dolphins, and porpoises) suborders produce characteristic sounds for communication and echolocation purposes.[1] Using hydrophones to record these sounds is a convenient non-invasive way of collecting information regarding social behavior, geographical distribution, etc. Much of this information can be automatically extracted from recordings using pattern-recognition algorithms, such as Gaussian mixture models (GMMs)[2] or hidden Markov models (HMMs),[3] which have been applied to a broad range of tasks ranging from audio to image processing.[4]

The main objective of this research was to automatically detect odontocetes and classify their sounds from underwater recordings. We focused specifically on different species of dolphins living off the northwest coast of Spain. The starting point was a database of some 81 h of recordings made during 2010 and 2011 under the LIFE+ INDEMARES project.[5] Using this database, we developed a GMM classifier to identify the two main kinds of sounds produced by dolphins, namely, whistles and pulses, and to distinguish these sounds from background noise. Whistles are narrow-band high-frequency signals, so it is difficult to model those using only frequency cepstral coefficients or other typical parameters found in many GMM-based audio and speech classifiers. To overcome this problem, we developed and computed two new parameters using the multiple signal classification (MUSIC) algorithm[6] and an unpredictability measure.[7] Tests on the database revealed substantial improvements in the detection and classification error rates when both these parameters were used. We also tested the new parameters for the more demanding task of whistle contour estimation and classification, with promising results.

The remainder of the article is organized as follows. Section II reviews recent research into the classification of cetacean sounds, Sec. III describes the databases used in our experiments, and Sec. IV describes the algorithms used to feed a classifier of different acoustic signals, which are themselves described in greater detail in Sec. V. Finally, Sec. VI describes experimental results and Sec. VII concludes the article.

II. STATE OF THE ART

Many research efforts have been directed to developing a system capable of classifying the sounds of different species of marine mammals. In some works,[3] cepstral coefficients have been used as features in HMMs and GMMs to classify different killer whale call types. Cepstral coefficients with GMMs have also been used to classify sounds for three different kinds of free-ranging odontocetes.[2] Furthermore, clustering of the first two cepstral coefficients and calculation of the slope for pulse sound onset have also been proposed as a way of estimating marine mammal numbers.[8]

Other authors have identified whistle contours in order to extract discriminant information capable of classifying different vocalizations. Datta and Sturtivant[9] encoded the contours with a quadratic equation to determine the identities of three dolphin groups using HMMs. Other algorithms have been proposed[10] whereby contours for different dolphin species were computed by applying particle filters (performing

[a]Author to whom correspondence should be addressed. Electronic mail: ppeso@gts.uvigo.es

Bayesian filtering) and adaptive polynomial prediction to a pre-processed spectrogram. A Kalman filter has also been proposed to trace the fundamental frequency of dolphin whistles.[11] The short-time Fourier transform has also been computed in some works[12] to classify different calls in a process that consists of extracting spectrogram peaks and connecting them using Bayesian inference.

Different features directly extracted from the contour (e.g., whistle duration, maximum and minimum frequency, etc.) were used by Díaz López[13] and Oswald *et al.*[14] to obtain information on the recorded whistles, with the Díaz López study proving that whistle type (defined only by these features) was correlated with the current behavior of the animal, and with the Oswald study identifying the species represented by the acoustic recordings using a MATLAB-based tool. Estimating whistle contour and pitch for vocal signals present similar difficulties in that both require locating a harmonic structure in the frequency domain. Some authors[15] have modified the discrete logarithmic Fourier transformation-pitch detection algorithm designed for human speech for application to killer whale vocalizations.

Classification performance greatly depends on the signal-to-noise ratio (SNR). A number of studies[11,16] have dealt with the problem of de-noising the spectrogram by applying image-processing methods. This approach enhances the marine mammal sounds and probably decreases the classification error rate (CER). Another alternative to removing noise is to decompose the signal in modes using the Hilbert–Huang transform.[17]

III. EXPERIMENTAL FRAMEWORK

The acoustic recordings used in this research were collected by CEMMA[18] under the Indemares Life 07/NAT/E/000732 project. The database consists of some 81 h of recordings captured during 2010 and 2011 in the Galician Bank and Avilés Canyon areas off the northwestern coast of Spain (Fig. 1). The audio files were recorded using 16-bit unsigned pulse-code modulation at 96 kHz.

The database also contains annotations regarding the cetacean species identified visually during the campaign. Using this information, researchers could identify and label the species represented in each recording. When no visual information was available, the label used was "non-identified delphinidus."

Species represented in the recordings were the common dolphin *Delphinus delphis* (DDE), the striped dolphin *Stenella coeruleoalba* (SCO), the common bottlenose dolphin *Tursiops truncatus* (TTR), the long-finned pilot whale *Globicephala melas* (GME), the sperm whale *Physeter macrocephalus*, and several unidentified cetaceans. Recordings for three species of dolphins (DDE, SCO, and TTR) were used for the automatic detection experiments, whereas the classification experiments also included recordings featuring the pilot whale (GME).

IV. METHODOLOGY

Below we describe a procedure to classify the two main sounds produced by odontocetes, namely, pulses and whistles.[1] Whistles are mainly used for communication and identification purposes, whereas pulses are used for social interaction (bursts) or echolocation (clicks). Odontocetes also produce other kind of noises (cracks, bangs, and pops), which were not addressed in our research.

Whistles, which have a narrow bandwidth and are short in duration (a few seconds at most), consist of omnidirectional tones that vary over time, sometimes presenting a strong harmonic structure. Figure 2 represents a sample extracted from our database.

Pulses are broadband signals with a directional pattern, extremely short in duration, and very frequently repeated, i.e., they have a low interpulse interval (no more than 3 ms for dolphins).[19] Pulses are of two kinds: echolocation clicks and burst pulses. The main difference is in the length of the interpulse interval, although other parameters, such as radiation patterns, also help to discriminate between the two sounds.[19] Figure 3 depicts a sample pulse train generated by an odontocete.

FIG. 1. (Color online) Map showing the area (offshore oval) where the data were recorded.

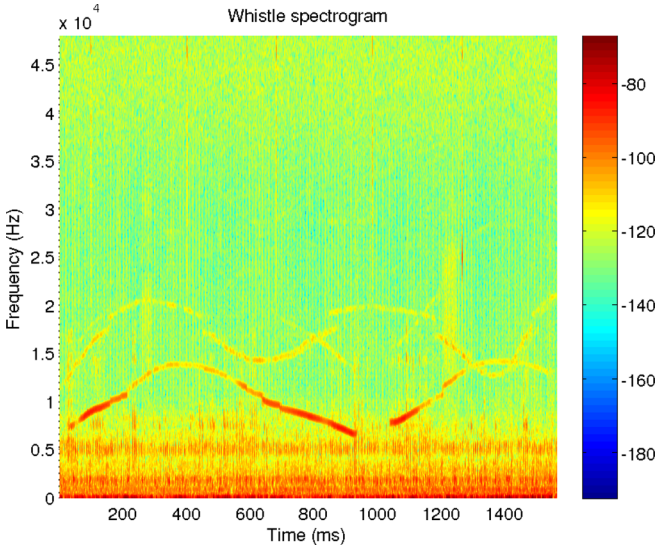

FIG. 2. (Color online) Spectrogram of a multiple whistle (*Tursiops truncatus*) sampled at 96 kHz.

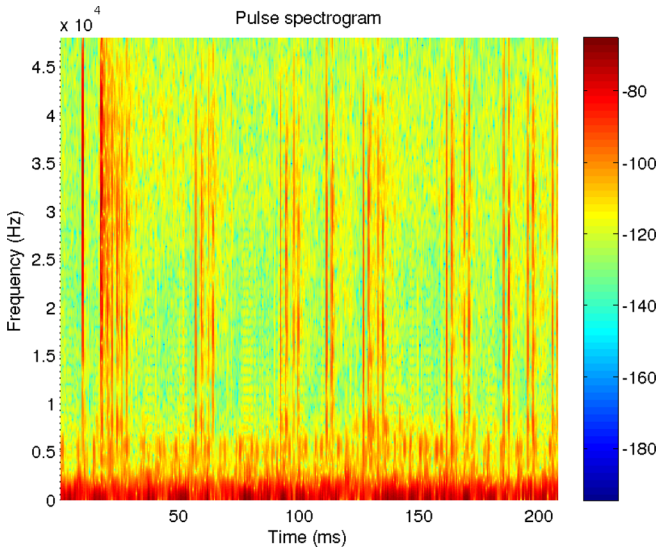

FIG. 3. (Color online) Spectrogram of a pulse train (*Tursiops truncatus*) sampled at 96 kHz.

Pulse and whistle characteristics vary between species and even between individuals of the same species. Whistles between species differ mainly in terms of frequency range and contour; pulses, meanwhile, have different rates, durations, frequency contents, and energy distributions.

Figure 4 depicts the main characteristics of the sound classification system. The first module, which parameterizes the input signal, extracts relevant features and removes redundant information from the raw signal. Three algorithms were used to extract features: (i) an algorithm based on cepstral transformation, broadly used for speech recognition purposes,[20] (ii) an algorithm that implements the unpredictability measure used in MP3 to estimate the predictability of the frequencies of a given frame using only information from the previous frame,[7] and (iii) Schmidts MUSIC algorithm,[6] which estimates the sinusoid frequencies in each frame. Although the first algorithm has been widely used in similar research, no references have been found in the literature as to use of the other two algorithms in this field.

To ensure correct feature extraction, the signal was first windowed by applying a hamming function to 2-ms frames with an overlap of 1 ms. Window size was selected taking into account the short duration of the pulses.

Below, we describe, in turn, the three techniques used to extract relevant features from the acoustic signal: cepstral coefficients, the unpredictability measure, and the MUSIC algorithm.

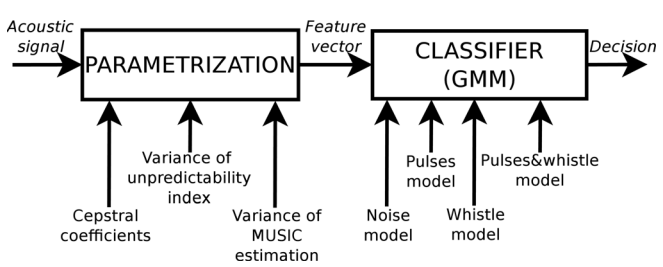

FIG. 4. Diagrammatic representation of the methodology.

A. Cepstral coefficients

The first features were extracted using the cepstral domain in a homomorphic transformation that separates the excitation signal from the periodic contributions, enabling cetacean detection by giving more uncorrelated coefficients compared to frequency magnitudes.

To compute the cepstral coefficients, a filter bank composed of N overlapping triangular filters was applied to the discrete Fourier transform of the frame. The log-energy of the output of each filter was then computed, obtaining a set of N log-energy spectral coefficients, m_j, $j = 1, 2, \ldots, N$. The M cepstral coefficients, c_i, were obtained applying an inverse discrete cosine transform using the following expression:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi_i}{N}(j-0.5)\right), \; i = 1, 2, \ldots, M. \quad (1)$$

In this case, we used $M = N = 12$; in other words, 12 cepstral coefficients were extracted from a bank of 12 filters equally spaced between 1 kHz and 30 kHz.

A typical speech recognition mechanism expands the feature vector by adding the first and second derivatives of the cepstral coefficients, with the aim of incorporating information about the time behavior of the parameters. We computed the derivatives using simple differences

$$d_i = \frac{c_{i+1} - c_{i-1}}{2}. \quad (2)$$

B. The unpredictability measure

Although cepstral coefficients are good at detecting pulses, they are not so useful in identifying whistles due to their narrowband nature, as whistle energy is easily confused with spectral noise captured by the triangular window. To overcome this problem we searched for a feature that would reveal the tonal nature of whistles even when embedded in noise and mixed in with pulses and bursts.

A good candidate was the tonality index,[7] commonly used for the MPEG I Layer3 audio coder to classify spectral peaks as tonal or noisy. A spectral sample is considered more tone-like if its amplitude and phase can be predicted using previous samples. The tonality index is therefore computed as the difference between the spectral sample and its predicted value. The assessment of this unpredictability measure for frequency bin, f, at frame, j, is defined as[7]

$$cw(f) = \frac{\sqrt{A+B}}{(R_j(f) + |\hat{R}_j(f)|)}, \quad (3)$$

$$A = (R_j(f)\cos(\phi_j(f)) - \hat{R}_j(f)\cos(\hat{\phi}_j(f)))^2,$$

$$B = (R_j(f)\sin(\phi_j(f)) - \hat{R}_j(f)\sin(\hat{\phi}_j(f)))^2,$$

where $R_j(f)$ and $\phi_j(f)$ represent the magnitude and phase of the component f, respectively. The predicted complex spectrum magnitude, $\hat{R}_j(f)$, and the predicted phase, $\hat{\phi}_j(f)$, were computed as follows:

$$\hat{R}_j(f) = 2R_{j-1}(f) - R_{j-2}(f),$$
$$\hat{\phi}_j(f) = \phi_{j-1}(f) + \phi_{j-2}(f). \tag{4}$$

This measure returned a vector with as many elements as frequency bins analyzed. A total of 1024 points were used. The variance of the vector was used to reduce it to just a single feature while retaining information on the presence of whistles.

Figure 5 shows the results for each step of the procedure as implemented. The top panel shows the spectrogram of the signal composed of merged noise, pulses, whistles, and pulses with whistles. Also represented is the value of $cw(f)$. Ideally, this tone-detection measure highlights only the whistles in the spectrogram. As can be observed, the method correctly detected the whistles in the last two segments. The next plotted step was a low-pass filter (in the horizontal direction) to remove outliers. Finally, the variance of the previous matrix was computed for all the frequencies in a given frame (i.e., in the vertical direction); this variance was used as a feature that increased for a whistle and otherwise decreased.

C. The MUSIC algorithm

Figure 4 shows a new parameter included in the feature vector after extraction using the MUSIC algorithm.

Although this algorithm was initially developed by Schmidt[6] to determine the parameters of signals received by an antenna, it quickly came to be adopted as a general spectrum estimation method. The MUSIC algorithm considers the incoming signal as a set of P sinusoids embedded in white Gaussian noise. In line with this premise, and assuming no correlation between incident signals and noise, the algorithm is capable of estimating the frequency and amplitude for each sinusoid. The main problem is that the number of sinusoids must either be previously known or otherwise inferred from the signal. Note that the underlying model was perfectly adequate for our purpose of detecting whistles embedded in noise.

MUSIC assumes that the signal $\mathbf{x}$ of length M is composed of a set of P sinusoids of frequencies ω_i plus a vector of M samples of white noise $\mathbf{w}$. The data model considered by the algorithm is as follows:

$$\mathbf{x} = \mathbf{aF} + \mathbf{w} = \sum_{i=0}^{P} a_i \mathbf{f_i} + \mathbf{w}, \tag{5}$$

where $\mathbf{a}$ is a vector containing the amplitudes and phases of the P sinusoid with components $a_i = |a_i|e^{j\phi_i}$, and $\mathbf{F}$ is a $P \times M$ matrix composed of P vectors of M samples of the signal frequencies $\mathbf{f_i} = [1\ e^{j\omega_i}\ \cdots\ e^{j\omega_i(M-1)}]$.

Under the assumption that incident signals and noise are uncorrelated, the P eigenvectors associated with the largest eigenvalues of the autocorrelation matrix of the received signal, $\mathbf{R_x}$, define the signal subspace. Meanwhile, the other $M - P$ eigenvectors describe the noise subspace, $\mathbf{E_N}$. The

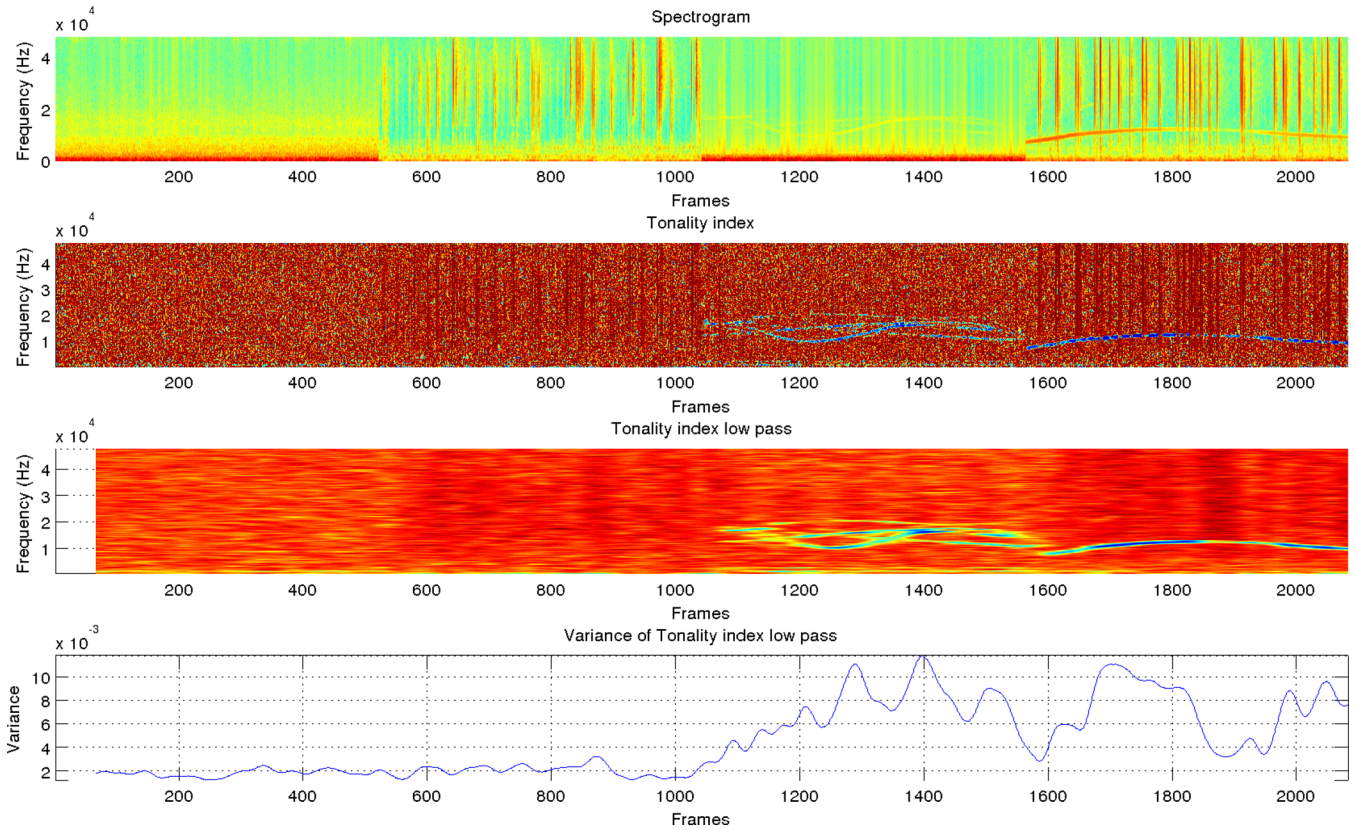

FIG. 5. (Color online) Unpredictability measure results for the four analyzed sound types (noise, pulses, whistles, and pulses with whistles). From top to bottom: the spectrogram for the four acoustic signals considered; $cw(f)$ for each frame; the result after applying an average filter (order 64) to the horizontal direction of $cw(f)$ (i.e., the unpredictability measure average for 64 consecutive frames for the same frequency bin); and the variance per frame, vertically computed from the previous representation.

Euclidean distance from vector $\mathbf{y}$ to the subspace $\mathbf{E_N}$ is given by $d^2 = \mathbf{y}^*\mathbf{E_N}\mathbf{E_N^*}\mathbf{y}$. Therefore, the maximum peaks in Eq. (6) provide the strongest frequencies, ω, in the signal

$$P_{MU}(\omega) = \frac{1}{\mathbf{a}^*(\omega)\mathbf{E_N}\mathbf{E_N^*}\mathbf{a}(\omega)}, \tag{6}$$

where "$*$" represents the Hermitian operation and $\mathbf{a}(\omega) = [1 \quad e^{jw} \quad \cdots \quad e^{jw(M-1)}]$.

Like the unpredictability measure, the MUSIC algorithm does not return a value that can be directly included in the feature vector. Hence, the predominant frequency (with $P = 8$) was computed for each frame and the feature was then extracted by computing the variance of these components for a window of 100 estimations. Figure 6 graphically depicts this method; the top panel shows the spectrogram for the noise, pulses, whistles, and pulses with whistles and the middle panel shows the predominant component between 1 kHz and 30 kHz for each frame. For any MUSIC-computed P frequencies outside these bounds, a random frequency was plotted, which caused greater variance in the next step. Estimated components that were close together in consecutive frames indicated a whistle. This variance was used as a feature that indicated the presence of a whistle if window variance decreased and vice versa (see Fig. 6).

This method yielded highly accurate estimates and performed well at a relatively low SNR, sometimes detecting whistles with spectral peaks of <1 dB above the background noise level. However, the transformations and the spectral peak search both required considerable computation time.

V. CLASSIFIER

Once the raw signal was parameterized, sounds were classified using features extracted by the first module. The second module (Fig. 4) decided the sound class (noise, pulses, whistles, or pulses with whistles) to which the feature vector belonged.

The classifier for the identification task was based on GMMs, composed of several Gaussian functions known as components. The probability estimate of a particular GMM, given data $\mathbf{X}$ (feature vectors), was computed as follows:

$$\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma},\pi) = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{X}_n|\boldsymbol{\mu}_k,\Sigma_k)\right\}, \tag{7}$$

where k represents the number of components. In this expression, which was used to fit the k multidimensional components to the data in order to create a representative model, the free parameters are as follows:

(1) Mean of the Gaussian, $\boldsymbol{\mu}_k$.
(2) Covariance matrix of the components, Σ_k.
(3) Mixing coefficient, π_k, i.e., the weight for each component. This parameter dimension is always 1 because it affects the whole Gaussian function. Also note that $\sum_{k=1}^{K} \pi_k = 1$.

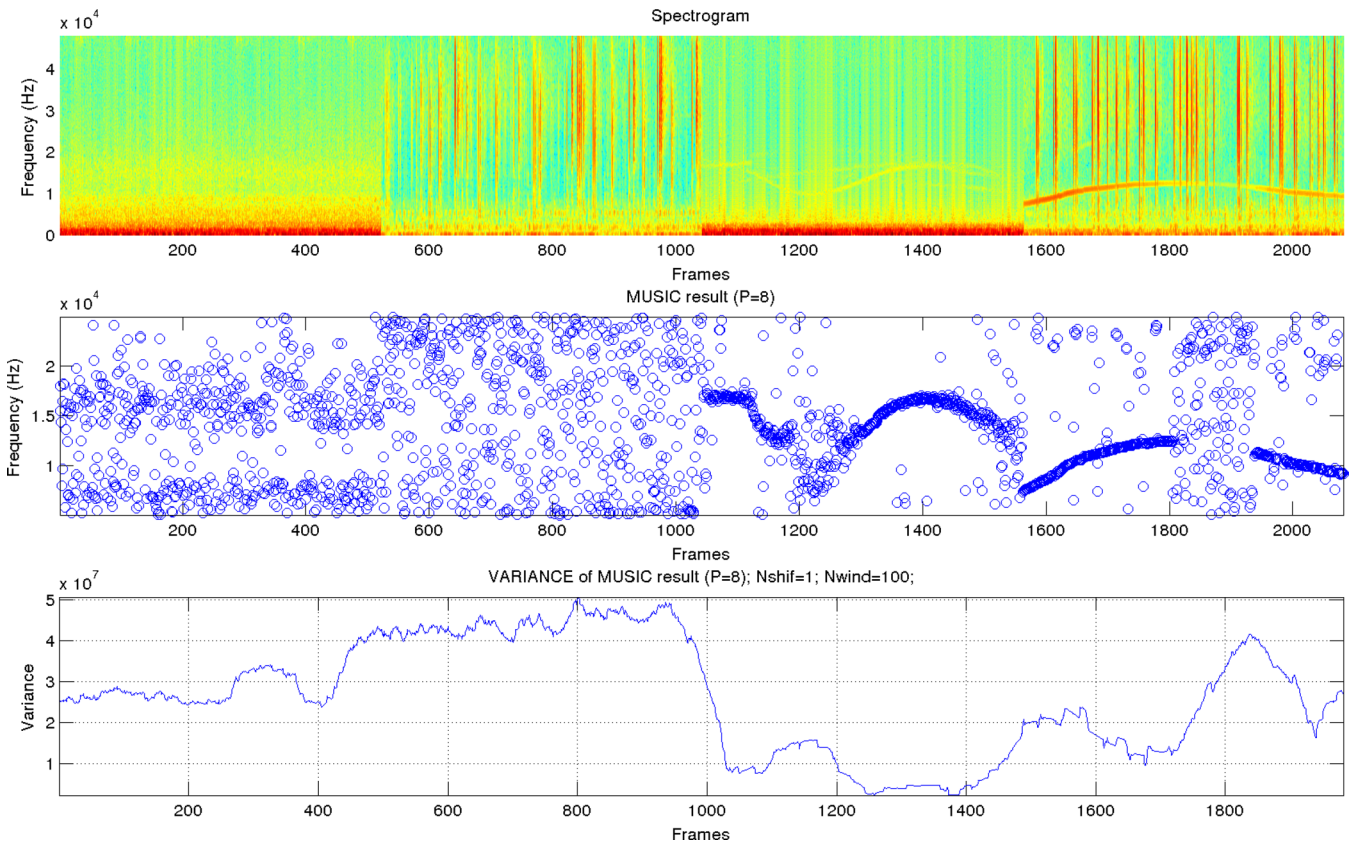

FIG. 6. (Color online) MUSIC algorithm results for the four analyzed sound types (noise, pulses, whistles, and pulses with whistles). From top to bottom: the spectrogram for the four different acoustic signals considered; the predominant component for each frame obtained with MUSIC; and the variance for the previous representation, plotted using a horizontal window of 100 points and a shift of 1 sample.

The expectation maximization algorithm was used to build these models ($\theta = \{\boldsymbol{\mu}, \Sigma, \pi\}$) adapted to the given data, $\mathbf{X}$. In this case, for the GMM, the expectation step is as follows:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{X}_n | \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{X}_n | \boldsymbol{\mu}_j, \Sigma_j)}, \quad (8)$$

where parameter $\gamma(z_{nk})$ indicates how much Gaussian k is responsible for point $\mathbf{X}_n$.

The maximization step maximizes the likelihood of the model reflected in Eq. (7) by using the parameter calculated in the previous step $\gamma(z_{nk})$ as follows:

(1) $\boldsymbol{\mu}_k^{\text{new}} = (1/\Sigma_{n=1}^{N} \gamma(z_{nk})) \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$,

(2) $\Sigma_k^{\text{new}} = (1/\Sigma_{n=1}^{N} \gamma(z_{nk})) \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^T$,

(3) $\pi_k^{\text{new}} = \Sigma_{n=1}^{N} \gamma(z_{nk})/N$.

This iteration was implemented until no significant improvement resulted or until the maximum number of iterations was reached.

A model was trained for each type of sound (noise, pulses, whistles, and pulses with whistles) or cetacean species depending on the experiment. The data, $\mathbf{X}$, used to train each model corresponded to the feature vectors extracted from the raw signal, tagged as the sound to be modeled.

Once models were trained for each sound, it was possible to classify the different acoustic signals. The strategy used was to maximize the posterior probability, $p(\theta|\mathbf{X})$, so as to determine which model was most likely to generate $\mathbf{X}$. The posterior probability computed according to the Bayesian theorem is as follows:

$$p(\theta|\mathbf{X}) = \frac{p(\mathbf{X}|\theta)p(\theta)}{p(\mathbf{X})}, \quad (9)$$

where $p(\mathbf{X})$ represents the prior probability of the observations and $p(\theta)$ represents the prior probability of the model.

In the maximization process, since it does not depend on the model, the denominator was excluded from this optimization as follows:

$$\theta_{\text{MAP}}(\mathbf{X}) = \arg\max_{\theta} p(\theta|\mathbf{X}) = \arg\max_{\theta} p(\mathbf{X}|\theta)p(\theta). \quad (10)$$

Therefore, the input signal was classified as the sound represented by the model, θ, that maximizes Eq. (10). Note that a uniform prior distribution was assumed, given that the prior probabilities of the models were unknown.

VI. RESULTS

Described, in turn, below are the results for three different experiments: evaluation of the sound classification system, testing of the unpredictability measure, and MUSIC algorithm as applied to whistle contour estimation and species classification using whistle contour estimates.

A. Sound classification

We selected 18×5-s segments for each sound (noise, pulse, whistle, and pulse with whistle), resulting in a 6-min database. The intention was to extract the four different sound types from each recording; however, since not all sound types were represented in some sessions, the missing sounds were extracted from other recordings.

Sound classification was evaluated using 72 recordings featuring the three dolphins, unidentified delphinids, and noise as follows: DDE (15), TTR (16), SCO (11), unidentified delphinids (11), and background noise (19). A leave-one-out cross-validation strategy was used for this evaluation; thus, one segment was used for evaluation and the remaining segments were used to train the models. The procedure was repeated for each segment and the decision as to sound type was taken segment by segment rather than frame by frame.

Four test configurations were implemented to evaluate performance using different sets of features. The first configuration used all the features (CC + UM + *MUSIC*); the second configuration considered only the cepstral coefficients (CC), the third configuration used the cepstral coefficients and the feature extracted from the unpredictability measure (CC + UM), and the fourth configuration consisted of the cepstral coefficients and the feature derived from the MUSIC algorithm (CC + *MUSIC*). Results are summarized in Tables I–III.

TABLE I. Confusion matrix for different sets of features after classifying the database. Each cell contains four numbers representing the number of times a given sound (actual sound) has been classified as a given sound type (predicted sound). "*" indicates the number of elements obtained using all the features; "#" indicates elements obtained using only the cepstral coefficients; "⋆" indicates elements obtained with the cepstral coefficients plus the unpredictability measure; and, finally, "▲" indicates the elements obtained with the cepstral coefficients plus the MUSIC algorithm. The remaining cells follow the same pattern.

		Predicted sound							
		Noise		Pulse		Whistles		Pulses with whistles	
Actual sound	Noise	16*	14#	1	3	1	1	0	0
		16⋆	15▲	2	1	0	2	0	0
	Pulses	4	3	13	14	0	0	1	1
		3	4	14	13	0	0	1	1
	Whistles	1	2	0	0	15	14	2	2
		1	2	0	0	14	13	3	3
	Pulses with whistles	0	0	1	4	2	1	15	13
		0	0	2	3	2	1	14	14

TABLE II. Classification error rates for different features applied to the database. CC represents cepstral coefficient features with first and second derivatives, UM represents the unpredictability measure feature, and *MUSIC* represents the MUSIC algorithm feature.

		Sounds			
		Noise	Pulses	Whistles	Pulses with whistles
Features	CC + UM + *MUSIC*	11.11%	27.78%	16.67%	16.67%
	CC	22.22%	22.22%	22.22%	27.78%
	CC + UM	11.11%	22.22%	22.22%	22.22%
	CC + *MUSIC*	16.67%	27.78%	27.78%	22.22%

TABLE III. Overall classification error and detection rates for different features applied to the database. CC represents cepstral coefficient features with first and second derivatives, UM represents the unpredictability measure feature, and *MUSIC* represents the MUSIC algorithm feature.

		Evaluation	
		Sound classification error rate	Presence detection rate
Features	CC + UM + *MUSIC*	18.06%	90.28%
	CC	23.61%	87.50%
	CC + UM	19.44%	91.67%
	CC + *MUSIC*	23.61%	87.50%

Table I shows the confusion matrix for the four tests with different combinations of features and performance levels for each sound type. The unpredictability measure (UM) parameter improved noise detection. Pulse detection decreased when all the features were used because *MUSIC* cannot detect a specific pulse as such, so pulse was classified as noise. For whistle detection, the best performance resulted from using all the features because UM properly classified the whistles that *MUSIC* failed to classify and vice versa. Thus, when both algorithms were used, whistles were correctly detected. The same happened for pulses with whistles, with use of all the features also resulting in the best performance.

The main reason for the failure to detect the remaining whistles was either because the SNR was low or the whistle did not appear continuously in the 5-s segments. Furthermore, *MUSIC* may not work properly when whistle slope is high or several whistles are present in the recording because (in both cases) variance would increase and the feature would be similar to that obtained for a non-whistle segment. The UM feature would perform better, nonetheless, because it would lead to higher variance (note that, with this method, unlike with *MUSIC*, high variance indicates whistle presence and vice versa).

Table II, which represents the CER for each sound assessed from the confusion matrix, indicates that the system with all the features is the best option for detecting all the sounds, except for pulses, where performance was least satisfactory for the *MUSIC* feature.

Table III, which shows the CER for the whole database, clearly indicates that the best performance was achieved using all the features. This table also shows the percentage of recordings in which the target mammals were correctly detected. The best set of features for these tasks corresponded to CC+UM, again, because of the impact of the *MUSIC* feature on pulse classification.

B. Contour estimation

Contour estimation aims to calculate the exact frequency of whistles in terms of two main aspects: the detection of all whistles in a signal; and the accuracy of the frequency estimate for each. This kind of knowledge can potentially be used to acquire further information about marine mammals (e.g., dolphin school size[21]). Below, we briefly evaluate the unpredictability measure and the MUSIC algorithm as used for these whistle modeling purposes.

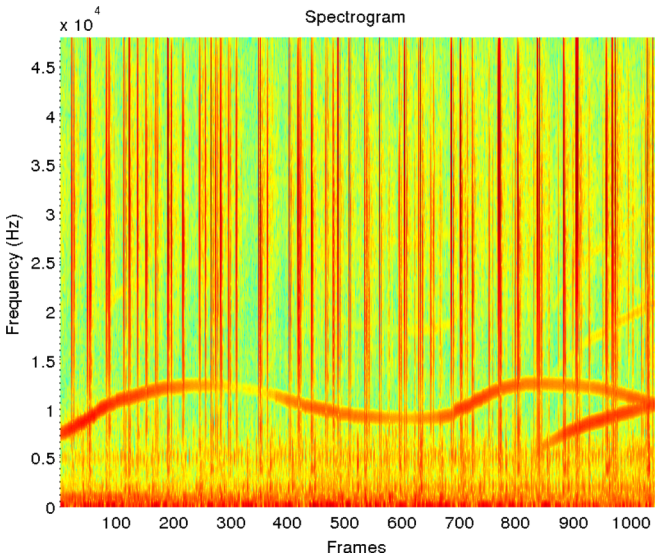

FIG. 7. (Color online) Spectrogram of a short simultaneous recording of pulses and whistles.

The unpredictability measure can be used to compute whistle frequency for each frame. The starting point is the result for the low-pass filter presented in Sec. IV B (third row in Fig. 5). Since the matrix contains contours represented with low values, in order to extract only this information, this matrix is binarized (using a threshold that might be trained). Figures 7 and 8 show the viability of this method for describing whistles.

The MUSIC algorithm can also be used to characterize whistle contours. Above, we described a tractable method to track these sound types whose first step consists of computing and storing the P components that lie within the studied frequency range, R, for each windowed frame. Once this assessment has been made for the entire audio signal, the points are connected, with two constraints: maximum

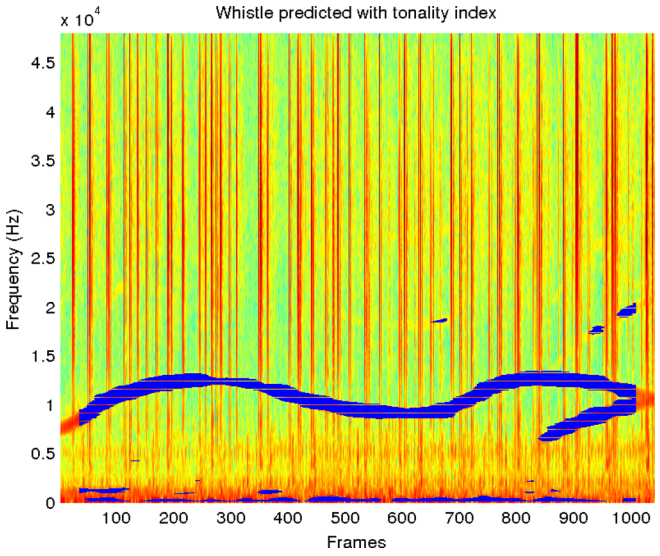

FIG. 8. (Color online) Whistle positions predicted using the unpredictability measure for the recording shown in Fig. 7.

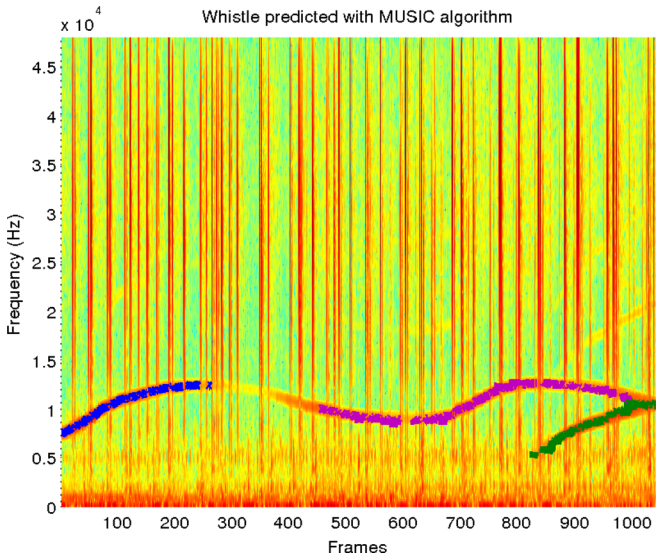

FIG. 9. (Color online) Whistle positions predicted using the MUSIC algorithm for the recording shown in Fig. 7.

whistle slope and maximum time between two consecutive components. The maximum whistle slope sets the maximum frequency difference for two components to be connected, while the maximum time establishes the maximum number of frames between two consecutive components. Figure 9 depicts this method for a specific signal [with $P = 10$ and $R = (5000–25000)$ Hz], with each color representing a different whistle.

Although the unpredictability measure lacks accuracy, unlike the MUSIC method, it searches the entire spectrum. The MUSIC approach seems to be more sensitive to a low SNR, as can be seen in Figs. 8 and 9: the MUSIC algorithm misses the path around frame 300, indicating that a combination of both methods could improve whistle contour estimation.

C. Species classification

This section concludes with some preliminary results pointing to the potential of the tonality index for cetacean classification. For this experiment, new training and test sets were used. The automatic whistle detector was first applied to the whole database and the resulting time boundaries were manually modified to correct incomplete whistles and remove detection errors. The information provided in the original database was used to label each whistle and so identify the cetacean species. The starting point was a set of some 3000 whistles, representing the three species of dolphins (DDE, TTR, and SCO) and the pilot whale (GME).

The segments were parameterized using the tonality index, and a slightly different procedure from that described in Sec. VI B. Thus, instead of binarizing the tonality index, we computed its maximum per frame, thereby obtaining a whistle contour estimation directly and avoiding the problem of identifying and separating overlapping whistles.

The analysis parameters were also modified to improve the continuity of the contours. We used a 3-ms Blackman–Harris window with 0.1 ms overlaps. These values improved continuity of the predictability measure from frame to frame, enabling the system to more easily follow whistle contours with high slopes. Finally, we applied a Savitzky–Golay filter to smooth the contours.

For this preliminary experiment, if at all possible, we wanted to discard noisy contours caused by overlapping whistles, strong harmonics, and weak recordings. For this purpose, we ordered the segments by variance and selected lower-value segments for each experiment. Table IV shows the material used for our experiments. The first row shows the total number and duration of the recordings initially available for each species. Note the great imbalance between the species, with DDE and GME having the most and fewest recordings, respectively. In the first experiment, we selected some 6 s of recordings for each dolphin species so as to match the amount of material available for GME. In the second experiment, we selected about 25 s for dolphin species and all the material available for GME. In both cases, the recordings contain isolated whistles and whistles mixed with pulses.

In interpreting these results, the limited number of GME recordings in the database should be taken into consideration. The material for this species was selected from three different recording sessions (i.e., from three different days) and so are likely to belong to three different individuals. For this reason, it may be assumed that the database contains only a limited subset of the repertoire of whistles for this species.

We trained four GMMs (one per species) using eight mixtures and diagonal covariance matrices. Figure 10 shows the probability density curve for each of the three components of the vector (contour frequency, and first and second derivatives) for the GMMs obtained using the second experiment defined in Table IV. It can be seen that the first and second derivatives seems to have null discriminative power, which might indicate that the whistle contour variation rate was similar for the four species. The top graph in the figure

TABLE IV. Database composition for experiments 1 and 2. Each cell shows the recording duration in seconds for each species and experiment. The number of whistles is shown in brackets.

		Cetacean species			
		DDE	SCO	TTR	GME
Database material	Total	271.7 s (1727)	94.0 s (845)	77.7 s (449)	5.8 s (14)
	Experiment 1	6.2 s (24)	6.3 s (14)	6.5 s (22)	5.8 s (14)
	Experiment 2	25.3 s (83)	25.0 s (60)	25.1 s (79)	5.8 s (14)

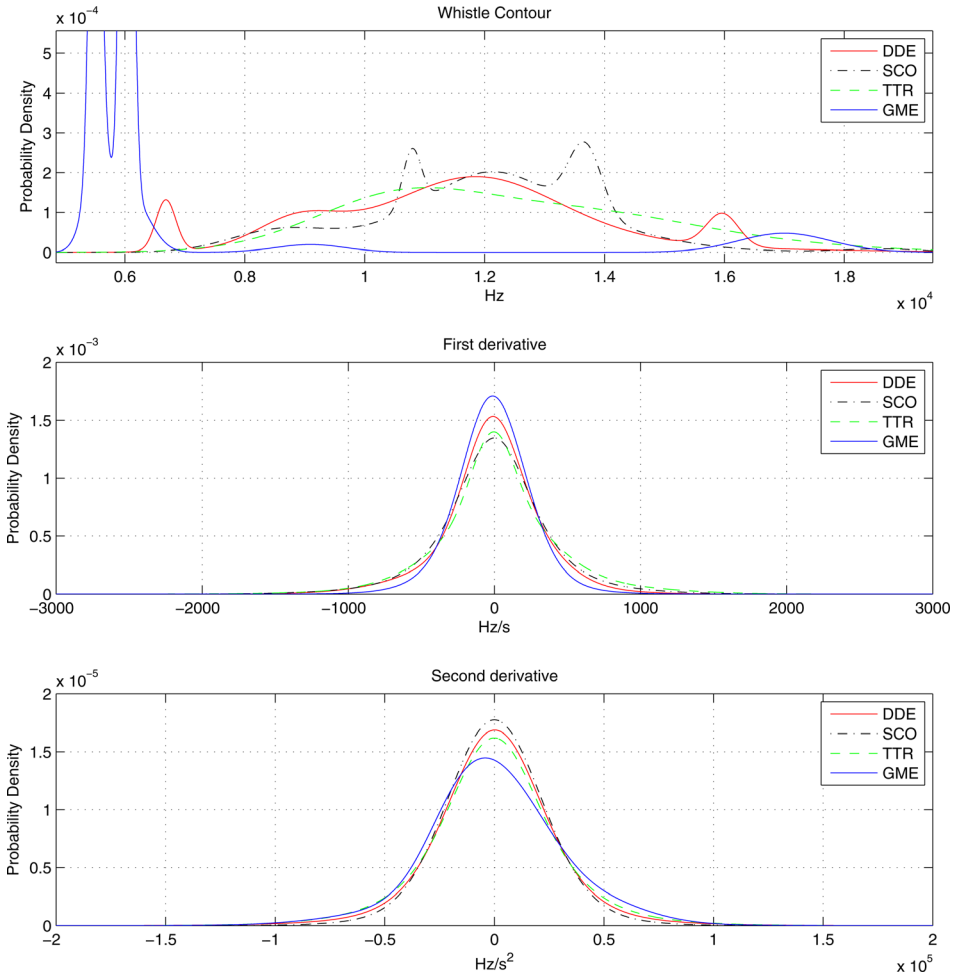

FIG. 10. (Color online) GMM components (probability density functions) for the whistle contours and the four cetacean species. From top to bottom: whistle contours, first derivative, and second derivative.

shows that, as expected, the whistles of the pilot whale were easily identifiable on the basis of its contour frequency. For the three dolphin species, however, there was a great overlap between the probability density functions.

We tested the classifier using the leave-one-out strategy described earlier. The resulting confusion matrices for the two experiments are shown in Tables V and VI. As can be seen, in both cases, GME was easily detected, for a classification rate of 73.3%. Of the dolphin species, only SCO appeared to be barely differentiable, with classification rates of 60% and 54% for the first and second experiments, respectively. TTR and DDE, meanwhile, seemed to be basically indistinguishable from each other using this approach. Note that in the second experiment, TTR and DDE were randomly classified as dolphin species, indicating that the classifier failed entirely to function.

Although the lack of GME recordings prevent us from drawing definitive conclusions, we consider that the tests presented here point to the tonality index as a good candidate for extracting and classifying cetacean whistles in noisy environments. The probability density functions obtained for the first and second derivatives, meanwhile, suggest that GMMs are not really suitable for this task. It is expected that a HMM with several states would be capable of modeling variations in the whistle contours, providing invaluable discriminating power.

TABLE V. Confusion matrix for experiment 1. Each cell shows the number and percentage of whistles for the species indicated in the rows classified as the species indicated in the columns. The diagonal of the matrix indicates the number and percentage of correctly classified whistles.

		Predicted species			
		DDE	SCO	TTR	GME
Actual species	DDE	11 (45.8%)	5 (20.8%)	8 (33.23%)	0 (0%)
	SCO	4 (26.7%)	9 (60.0%)	1 (6.7%)	1 (6.7%)
	TTR	8 (34.8%)	4 (17.4%)	10 (43.5%)	1 (4.3%)
	GME	3 (20.0%)	0 (0%)	1 (6.7%)	11 (73.3%)

TABLE VI. Confusion matrix for experiment 2. Each cell shows the number and percentage of whistles for the species indicated in the rows classified as the species indicated in the columns. The diagonal of the matrix indicates the number and percentage of correctly classified whistles.

		Predicted species			
		DDE	SCO	TTR	GME
Actual species	DDE	26 (30.9%)	31 (36.9%)	26 (30.9%)	1 (1.2%)
	SCO	10 (16.4%)	33 (54.1%)	17 (27.9%)	1 (1.6%)
	TTR	27 (33.7%)	23 (28.7%)	29 (36.2%)	1 (1.25%)
	GME	2 (13.3%)	0 (0%)	2 (13.3%)	11 (73.3%)

VII. CONCLUSION

We have described a system for detecting cetacean species and classifying their sounds as pulses, whistles, or pulses with whistles. Noise was included as a fourth sound as a way of detecting cetacean presence. We also assessed the performance of two novel approaches to whistle contour estimation.

Detection of pulses with whistles using cepstral coefficients alone presented some issues that we addressed by including two novel features extracted from the unpredictability measure and the MUSIC algorithm, thereby improving the whistle detection rate and overall system performance. The system also proved to be highly accurate in detecting cetacean presence, although the MUSIC algorithm tended to confuse some pulses with noise.

No reference has been encountered in the relevant literature to the features extracted from the unpredictability measure and the MUSIC algorithm, yet it would seem that these parameters show some promise in terms of achieved satisfactory results for contour whistle estimation and species classification. Nevertheless, they need to be further evaluated using various metrics and in experiments using segments with several whistles.

A future research line is evaluation of the system over entire recordings as a more realistic environment. The system could also include a metric to analyze captured pulses and whistles in order to determine dolphin school sizes. Another interesting line of research concerns the use of signature whistles to detect and locate specific individuals.

ACKNOWLEDGMENTS

The authors would like to thank CEMMA (Coordinadora para o Estudo dos Mamíferos Mariños) for making underwater recordings and information on the studied marine mammals and their behavior available to the researchers. This work has been supported by the European Regional Development Fund, the Galician Regional Government (Grant Nos. CN2011/019 and CN2012/160), and a Campus do Mar scholarship for post-graduate master studies.

[1]W. W. L. Au, A. N. Popper, and R. R. Fay, "Hearing by whales and dolphins," in *Springer Handbook of Auditory Research* (Springer, New York, 2000), pp. 1–485.

[2]M. A. Roch, M. S. Soldevilla, J. C. Burtenshaw, E. E. Henderson, and J. A. Hildebrand, "Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California," J. Acoust. Soc. Am. **121**(3), 1737–1748 (2007).

[3]J. C. Brown and P. Smaragdis, "Hidden Markov and Gaussian mixture models for automatic call classification," J. Acoust. Soc. Am. **125**(6), EL221–EL224 (2009).

[4]R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. (Wiley, New York, 2001), pp. 1–654.

[5]http://www.indemares.es/index.php?Itemid=59&id= 4&option=com_content&lang=en (Last viewed 3/12/13).

[6]R. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986).

[7]J. J. Thiagarajan and A. Spanias, "Analysis of the MPEG-1 Layer III (MP3) algorithm using MATLAB," in *Synthesis Lectures on Algorithms and Software in Engineering* (Morgan and Claypool, San Rafael, 2011), pp. 1–130.

[8]X. C. Halkias and D. P. W. Ellis, "Estimating the number of marine mammals using recordings of clicks from one microphone," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France (May 14–19, 2006), Vol. 5, pp. 769–772.

[9]S. Datta and C. Sturtivant, "Dolphin whistle classification for determining group identities," Signal Process. **82**(2), 251–258 (2002).

[10]M. A. Roch, T. S. Brandes, B. Patel, Y. Barkley, S. Baumann-Pickering, and M. S. Soldevilla, "Automated extraction of odontocete whistle contours," J. Acoust. Soc. Am. **130**(4), 2212–2223 (2011).

[11]A. Mallawaarachchi, S. H. Ong, M. Chitre, and E. Taylor, "Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles," J. Acoust. Soc. Am. **124**(2), 1159–1170 (2008).

[12]X. C. Halkias and D. P.W. Ellis, "Call detection and extraction using Bayesian inference," Appl. Acoust. **67**(11-12), 1164–1174 (2006).

[13]B. Díaz López, "Whistle characteristics in free-ranging bottlenose dolphins (*Tursiops truncatus*) in the Mediterranean Sea: Influence of behaviour," Mamm. Biol. - Z. Säugetierkunde **76**(2), 180–189 (2011).

[14]J. N. Oswald, S. Rankin, J. Barlow, and M. O. Lammers, "A tool for real-time acoustic species identification of delphinid whistles," J. Acoust. Soc. Am. **122**(1), 587–595 (2007).

[15]A. D. Shapiro and C. Wang, "A versatile pitch tracking algorithm: From human speech to killer whale vocalizations," J. Acoust. Soc. Am. **126**(1), 451–459 (2009).

[16]C. R. Sturtivant and S. Datta, "Techniques to isolate dolphin whistles and other tonal sounds from background noise," Acoust. Lett. **18**(10), 189–193 (1995).

[17]O. Adam, "Segmentation of killer whale vocalizations using the Hilbert–Huang transform," EURASIP J. Adv. Signal Process. **2008**(162), 1–10 (2008).

[18]http://www.cemma.org (Last viewed 3/12/13).

[19]W. W. L. Au and M. C. Hastings, *Principles of Marine Bioacoustics*, 1st ed. (Springer, New York, 2008), pp. 1–679.

[20]S. Molau, M. Pitz, R. Schlüter, and H. Ney, "Computing Mel-frequency cepstral coefficients on the power spectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, UT (May 2001), Vol 1, pp. 73–76.

[21]S. M. Van Parijs, J. Smith, and P. J. Corkeron, "Using calls to estimate the abundance of inshore dolphins: A case study with Pacific humpback dolphins *Sousa chinensis*," J. Appl. Ecol. **39**, 853–864 (2002).