Call recognition and individual identification of fish vocalizations based on automatic speech recognition: An example…

Article *in* The Journal of the Acoustical Society of America · December 2015

DOI: 10.1121/1.4936858

4 authors, including:

P.J. Fonseca
University of Lisbon
95 PUBLICATIONS **835** CITATIONS
SEE PROFILE

M. Clara P Amorim
ISPA Instituto Universitário
88 PUBLICATIONS **1,360** CITATIONS
SEE PROFILE

Carlos Teixeira
University of Lisbon
29 PUBLICATIONS **153** CITATIONS
SEE PROFILE

Call recognition and individual identification of fish vocalizations based on automatic speech recognition: An example with the Lusitanian toadfish

Manuel Vieira and Paulo J. Fonseca
Departamento de Biologia Animal and cE3c - Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de Lisboa, Bloco C2. Campo Grande, 1749-016 Lisboa, Portugal

M. Clara P. Amorim
MARE—Marine and Environmental Sciences Centre, ISPA—Instituto Universitário, Rua Jardim do Tabaco 34, 1149-041 Lisboa, Portugal

Carlos J. C. Teixeira[a)]
Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Bloco C6. Campo Grande, 1749-016 Lisboa, Portugal

(Received 16 January 2015; revised 29 October 2015; accepted 30 October 2015; published online 30 December 2015)

The study of acoustic communication in animals often requires not only the recognition of species specific acoustic signals but also the identification of individual subjects, all in a complex acoustic background. Moreover, when very long recordings are to be analyzed, automatic recognition and identification processes are invaluable tools to extract the relevant biological information. A pattern recognition methodology based on hidden Markov models is presented inspired by successful results obtained in the most widely known and complex acoustical communication signal: human speech. This methodology was applied here for the first time to the detection and recognition of fish acoustic signals, specifically in a stream of round-the-clock recordings of Lusitanian toadfish (*Halobatrachus didactylus*) in their natural estuarine habitat. The results show that this methodology is able not only to detect the mating sounds (boatwhistles) but also to identify individual male toadfish, reaching an identification rate of ca. 95%. Moreover this method also proved to be a powerful tool to assess signal durations in large data sets. However, the system failed in recognizing other sound types. © *2015 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4936858]

[ANP] Pages: 3941–3950

I. INTRODUCTION

Many species communicate through acoustic signals that can fulfill several functions from mediating agonistic interactions to reproductive activities. Being able to monitor extensive acoustic recordings in nature can be a valuable tool to monitor activity and distribution of important species (e.g., Küsel *et al.*, 2011).

Several approaches have been reported to study extensive bioacoustic recordings. The simplest and most common are automatic detection methods that make use of, for example, energy thresholds or a matched filter to locate the vocalizations, usually followed by common multivariate statistical analysis procedures to categorize the vocalizations (e.g., discriminant function analysis or linear discriminant analysis). With the advances of automatic speech recognition models and techniques in the past few decades, the recognition of sound patterns has become increasingly faster, accurate, and robust. Robust methods using machine learning, such as Gaussian mixture models (GMMs; Reynolds and Rose, 1995), artificial neural networks (ANN; Lippmann, 1988; Yu and Oh, 1997), and hidden Markov models (HMMs; Baker, 1975; Jelinek, 1976; Jelinek *et al.*, 1975; Rabiner, 1989; Young and Bloothooft, 1997), have been reported to successfully recognize and classify human and other animals' vocalizations. Table I refers to application examples with sounds of different animals such as insects, birds, amphibians, and mammals. HMM is the most used statistical model for automatic speech recognition systems (Dahl *et al.*, 2012).

HMM based approaches (Rabiner, 1989) successfully allow identification of species and individuals, examination of vocal repertoires, and classification of vocalizations according to social context or behaviour (Table I). HMMs can be used to statistically model both temporal and spectral variations of vocalizations through robust algorithms allowing optimization of relevant mathematical criteria. These approaches are capable to deal with extensive recordings allowing recognition and classification of animal vocalizations.

Fish often use acoustic signals during mating and territorial defense and are probably the largest sound producing vertebrate group (Ladich, 2004), but to the best of our knowledge, no automatic detection machine learning HMM-based application to fish sounds has been attempted (but see an application of a GMM approach to detect and count shad splashing spawning sounds; Diep *et al.*, 2013). The family

[a)]Electronic mail: cjteixeira@Ciencias.ULisboa.Pt

TABLE I. Examples of automated recognition applications in bioacoustics. ANN, artificial neural network; GMM, generalized method of moments; LFCC, linear frequency cepstral coefficients; LDA, linear discriminant analysis; DT, decision tree; SVM, support vector machine; DTW, dynamic time warping; MFC, Mel-frequency cepstral; LPC, linear prediction cepstral; S, species recognition, I, individual identification; C, call type recognition; TDSC, time domain signal coding; SCF, spectrogram correlator filter; WPD, wavelet packet decomposition.

Class		Objective	System	Feature	Reference
Insects		S	ANN	TDSC	Chesmore (2008)
	Orthoptera	S	ANN	TDSC	Chesmore (2001)
		S	ANN	TDSC	Chesmore and Ohya (2004)
	Cicadas	S	GMM	LFCC	Potamitis (2007)
Birds		S	ANN	LPC	McIlraith and Card (1995)
		S	ANN		Mills (1995)
		C	DTW		Anderson *et al.* (1996)
		C	LDA	MFC, LPC	Lee *et al.* (2006)
		S	HMM	[a]	Chou *et al.* (2007)
		S,C	ANN	MFC	Chou *et al.* (2008)
		S	HMM	MFC	Trifa *et al.* (2008)
		S	LDA,DT,SVM	[a]	Acevedo and Corrada-Bravo (2009)
		C	HMM	LPC	Chu and Blumstein (2011)
	Corncrake	I	ANN	[a]	Terry and McGregor (2002)
	Norwegian Ortolan Bunting	I, C	HMM	MFC	Trawicki (2005)
Amphibians		S,I	ANN	WPD	Yen and Fu (2001)
		S	LDA,DT,SVM	[a]	Acevedo and Corrada-Bravo (2009)
Mammals	Cetaceans	S	ANN	SCF	Potter *et al.* (1994)
		C	ANN	[b]	Murray *et al.* (1998)
		C	ANN	[c]	van der Schaar *et al.* (2007)
		C	HMM	LPC, MFC	Pace *et al.* (2012)
	Deer	I	ANN	MFC	Reby *et al.* (2006)
	Bats	S	ANN	[a]	Parsons and Jones (2000)
		S	ANN	[a]	Parsons (2001)
	Pigs (stress calls)	C	ANN	LPC	Schön *et al.* (2001)
	Sea lions ♀	I	ANN	[d]	Campbell *et al.* (2000)
	Elephants	I, C	HMM	MFC, PLP	Clemins *et al.* (2005)
	Cows	C	HMM	MFC	Jahns (2008)
	Primates	C	ANN	LPC	Pozzi *et al.* (2010)

[a]Vector composed of call variables;
[b]each vocalization was characterized by its simultaneous modulations in duty cycle and peak frequency;
[c]features were selected using a local discriminant basis;
[d]each call was represented by an average logarithmic spectrum on the back propagation network input layer.

Batrachoididae includes several species that have become good models for acoustic communication studies, such as the Lusitanian toadfish, *Halobatrachus didactylus* (Bass and McKibben, 2003; Vasconcelos *et al.*, 2012). This species is highly vocal and has an unusually large acoustic signal repertoire for fish that includes boatwhistles, croaks, double croaks, long grunt trains, grunts, and other less frequent sound combinations (see Amorim *et al.*, 2008, for details of the vocalizations). During the breeding season in Portugal, the species can be found in estuarine shallow waters, often presenting high turbidity, where breeding males occupy nests under rocks and produce boatwhistles to attract females (Vasconcelos *et al.*, 2012). The advertisement boatwhistle, the most frequent sound in this species, is a highly stereotyped low-frequency signal with a duration ranging from 400 to 1200 ms and a dominant frequency between ca. 50 and 200 Hz (Amorim and Vasconcelos, 2008; Amorim *et al.*, 2008). Interestingly, the boatwhistle presents inter-individual differences during short periods of time ($<$ 10 min; Amorim and Vasconcelos, 2008; Amorim *et al.*, 2011), allowing the recognition of individuals based on sounds.

We implemented call recognition and individual identification methods for the Lusitanian toadfish using HMMs. This allowed estimation of the duration and overlap of a time-running window and choice of features for the signal processing module. Additionally, the HMM topology and grammar were redesigned. These adaptations were specific for *Halobatrachus didactylus* but can be adjusted to other species. We chose this fish model because of the richness of its vocal repertoire, its sedentarism during the breeding season, and the ease of access to its breeding site. The possibility to analyse multiple round-the-clock recordings will allow inferring relevant ecological information such as vocal rhythms, acoustic social interactions and variability, possible effects of environmental parameters, anthropogenic noise, etc.

II. METHODS

A. Data collection

We recorded the vocalizations of adult territorial males during the breeding season (May to July 2012). The males

spontaneously occupied concrete artificial hemicylindrical nests, capped at one end, which we deployed in the Tagus estuary (Air Force Base 6, Montijo, Portugal; 38°42′N, 8°58′W). These nests, positioned at 2 m from each other in a row, had the entrance covered with a stainless steel net with an opening large enough to allow females or small prey (e.g., crabs) to enter the nest but prevented the larger territorial males from escaping. The males' vocalizations were recorded with custom-made hydrophones (Fonseca and Maia Alves, 2011) placed next to each experimental nest in mid-lateral position and about 10 cm above the substrate. The sounds produced by males occupying adjacent nests arrived much attenuated relative to the nest-holder vocalizations ensuring unequivocal individual signal identity throughout the recordings. The acoustic signals should be sampled in a limited bandwidth with a spectral range corresponding to the receivers' hearing capabilities, which in the Lusitanian toadfish does not go beyond 1 kHz (Vasconcelos *et al.*, 2007). Limiting the bandwidth may allow reducing extraneous noise and hence to improve signal-to-noise ratio. The signal from each hydrophone was recorded to a 16 channel stand-alone data logger (Measurement Computing Corporation LGR-5325, Norton, VA, 16 bit resolution, 4 kHz sampling rate, two times above the required Nyquist sampling rate for the target bandwidth).

The data set consisted of ca. 12 day round-the-clock simultaneous recordings of 16 nest-holders. The boatwhistles recorded in each channel (nest) were manually selected, classified, and cut with the aid of a matched filter function available in ISHMAEL 1.0 (Mellinger, 2002). Each individualized vocalization, delimited by 0.5 s of background noise, was stored in a ca. 2 s separate file. Only 13 males produced sounds in a total of 14 795 boat whistles, 23 croaks, 24 double croaks, and 77 grunts. Note that long grunt trains, the second most common vocalization (Amorim *et al.*, 2008) have lower energy and were not considered in this study. Due to the scarcity of croaks, double croaks, and grunts, we provided the training data set with at least six of each of these signal types, obtained from previous recordings (July to September 2002; Amorim *et al.*, 2006).

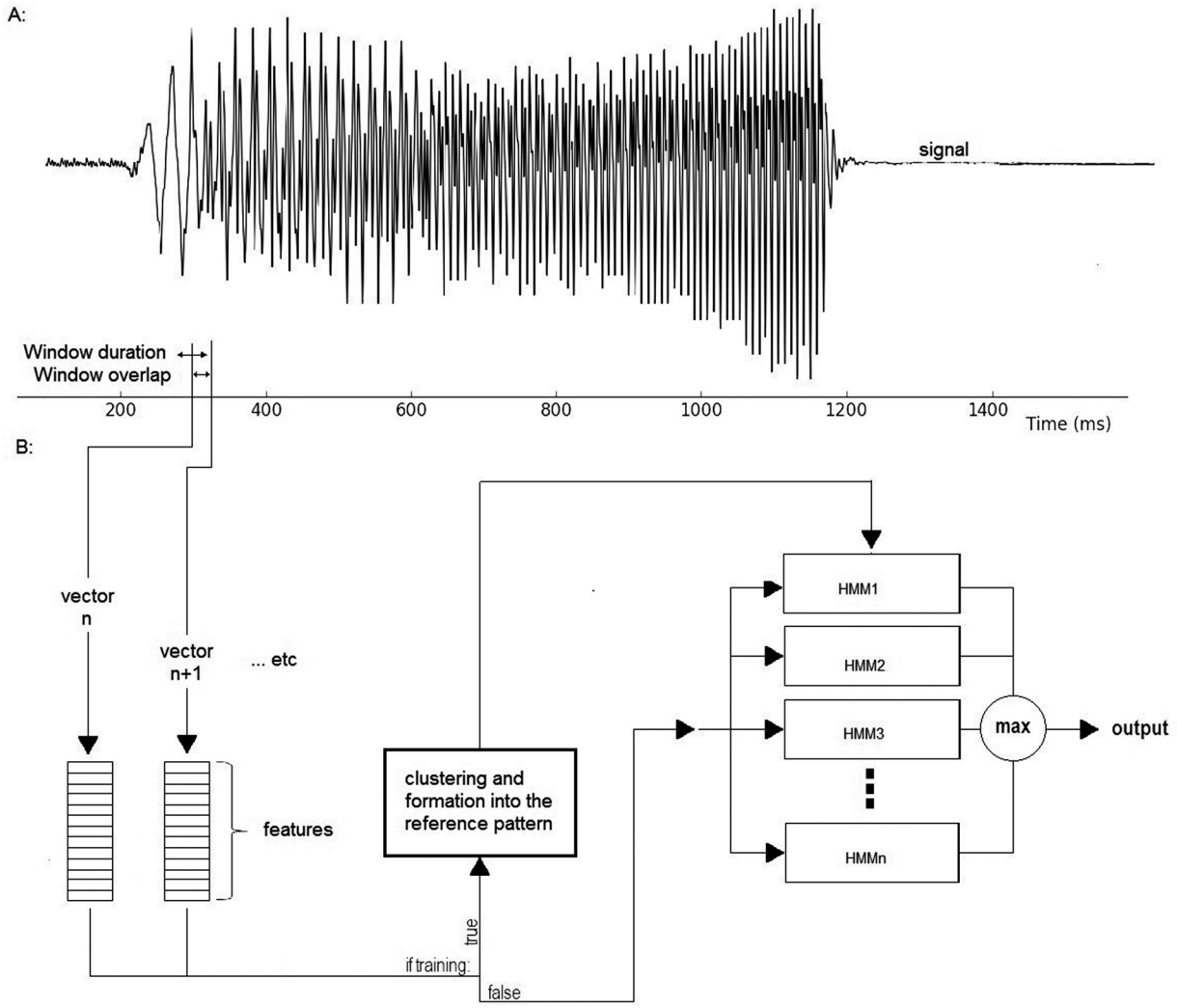

FIG. 1. Workflow of the HMM recognition system using the HMM ToolKit (HTK, diagram based on Young *et al.*, 2006). (A) The signal represents an oscillogram of a boatwhistle. The window duration, that defines the signal segmentation and the window overlap (target size) are represented. (B) The features' vector that allow computing the statistical parameters of the HMMs used in the recognition system.

TABLE II. Signal extraction features collected for every elementary segment.

Description	Reference
Cepstrum is the inverse Fourier transform of the logarithm of the signal spectrum. Applied to discrete data it is calculated using the discrete cosine tranform. $$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \log |X(k)| e^{j2\pi kn/N}$$	Oppenheim (1974), Young *et al*. (2006).
Mel-frequency cepstrum (MFC) is calculated using the discrete cosine transform on the mel scale of frequency. This is a series of tones perceived by a listener as being equally spaced in the frequency domain.	Bridle and Brown (1974)
Perceptual linear predictive (PLP) is an analysis method that takes into consideration perception both in frequency domain and amplitude.	Hermansky (1990)
Delta (Δ) calculated with $\Theta = 2$ for any coefficient c(i): $$\Delta_c(i) = \frac{\sum_{n=1}^{\Theta} n[c(i+n) - c(i-n)]}{2\sum_{n=1}^{\Theta} n^2},$$	Young *et al*. (2006)
Acelleration is a difference of coefficients delta. This difference is calculated with the expression used for delta.	Young *et al*. (2006)

B. Pattern recognition

Figure 1 summarizes the main stages of the signal recognition system, which include the signal processing for feature extraction and the alignment of the obtained feature vectors with several previously trained HMM models.

1. Signal processing

The first stage in the signal processing (Fig. 1) segments the waveform signal, according to a predefined window duration, into a sequence of elementary segments. This window should be longer than a cycle of the lower relevant frequency but short enough to provide temporal resolution while also assuring stable properties. Note that the lowest fundamental frequency of the different toadfish sound types is approximately 50 Hz (boatwhistles), corresponding to a pulse period of 20 ms (Amorim *et al*., 2008). Based on this value, we tested several window durations and selected 32 ms for the elementary segment of the toadfish vocalizations as it maximized correct signal classification. A 50% window overlap was used to avoid losing information on the transition between two consecutive elementary segments (O'Shaughnessy, 1987).

To investigate what signal features allowed better discrimination of vocalizations, we tested several combinations of relevant features: energy, cepstrum, Mel-frequency cepstral (MFC), perceptual linear prediction (PLP), delta, and acceleration coefficients (Table II), also used in previous bioacoustics studies (Table I; see Young *et al*., 2006, for details in these signal features).

The system was optimised by testing different frequency bandwidths adjusted to the spectrum of the vocalizations recorded in the natural habitat. These tests considered different low (0, 10, 20, 30, 40, and 50 Hz) and high (300, 500, 700, 1000, 2000 Hz) frequency cut-offs.

2. The HMM structure time alignment

A Markov model characterizes data in a sequence of states (in Fig. 2), each with a probability depending only upon the previous state (the Markov property). The transition between any two states S_i and S_j is governed by a discrete

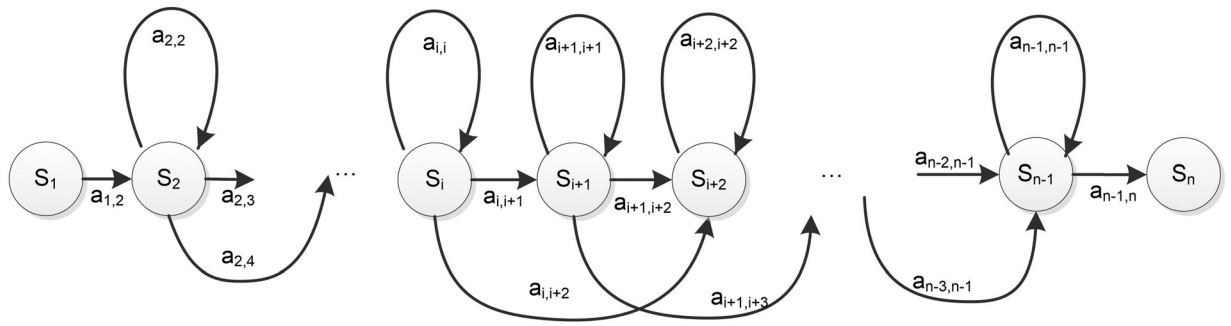

FIG. 2. The state model topology used for the adopted hidden Markov model is presented where states are represented by S_i, and the transitions are labelled with the corresponding transitions probabilities $a_{i,j}$. S_1, and S_n are the initial and final states, respectively.

probability $a_{i,j}$. In a HMM, the underlying state sequence is hidden and only the considered features (e.g., cepstrum, MFC, etc., see Sec. II A in the preceding text) are observed. Thus each state may be described via a probability density function (Gaussian mixture density) for the chosen features. The best set of state transition probabilities (transition matrix) is obtained by training the models with real data. The output of this training phase is a recognition system accounting not only for features of elementary signal segments and their variability but also for their sequence, thus establishing robust recognition of sounds (Baker, 1975; Jelinek, 1976; Jelinek *et al.*, 1975; Rabiner, 1989; Young and Bloothooft, 1997).

The use of Markov models for classification of acoustic signals in the time domain is naturally associated with linear topologies where each state has a transition ($a_{i,i+1}$ in Fig. 2) to the following one but not backward, thus imposing a time structure. This kind of topologies often also provide additional transitions ($a_{i,i}$ in Fig. 2) to the same state to recognise stable parts of the signal that are longer than the elementary segment. Following a similar reasoning, for segments significantly shorter than that minimum, an additional transition is often included between states S_i and S_{i+2} (Fig. 2). In the present work, we used a linear topology with additional transitions to the same state and to two states ahead. Notice that self-transitions are meaningless for initial and final states because they only serve as signal boundary markers (S_1 and S_n in Fig. 2).

Comparing each toadfish vocalization to a human spoken word, each state in a HMM can then be compared to a human language phoneme. Each word, as each phoneme, has an average expected duration that is directly related to the number of states. For example, a human phoneme is usually modelled by three states (McDermott *et al.*, 1990). However, because we do not have a phoneme set for the toadfish, we assumed that the number of states should be equal to the number of different consecutive stable parts of the signal as recognised by the HMM. We considered 14 states for the boatwhistles, 5 for the croak, 7 for the double croak, and 14 for the grunt train models. We further defined extra models with five states for modelling background noise and all non-biological patterns with high energy (e.g., small wave splash).

For each sound type, a representative subset of samples was used to train the HMMs. The transition probabilities and the elementary segment probability densities of each state were estimated with the Baum–Welch algorithm (Baum *et al.*, 1970).

In the recognition phase, each vocalization was matched against the estimated HMM for each sound type. This was achieved by using a Viterbi algorithm (Forney, 1973) that produced a likelihood measure for each HMM. The vocalization was assigned to a sound type considering the HMM with the highest likelihood. In addition, the time interval from the initial to the final states of each HMM provided an estimate of the sound duration (Sec. III C).

For computations, we used the HMM Toolkit (HTK, University of Cambridge, UK), which is a group of modules written in C to create automatic recognition systems for human speech analyses (Young *et al.*, 2006).

3. Automatic recognition systems

Two automatic HMM-based systems were prepared, one to identify boatwhistles of one individual among several vocalizing toadfish and another to identify the different sound types.

The individual identification system was based on HMMs for each individual fish trained with its own boatwhistles. We used a training set with 3–40 boatwhistles randomly selected from each fish to establish the minimum size of the training data set to produce an adequate recognition system. Nine signals were sufficient to reach more than 90% of correct identification rate (see Fig. 3). In a first set of trials, nine boatwhistles were randomly taken from the whole recordings, while in the second, only from the first 50 boatwhistles produced by each fish. The system was tested with the remaining boatwhistles. To take full advantage from the available data, a resampling method was used based on a random subsampling validation (Efron, 1981). Hence these trials were repeated 100 times with different training sets. These were again repeated for each feature set mentioned in Sec. II B. To subsequently analyse the recordings, we used the feature set that presented the highest individual identification scores.

The sound type identification system was trained for each signal type using sounds from all fish. In addition sounds obtained in previous seasons and available from our sound archive (Amorim *et al.*, 2006) were also used. A total of 14 795 boat whistles, 23 croaks, 24 double croaks, and 77 grunts were used. From these, nine sounds (see preceding text) were randomly resampled to include in the training set for each vocalization (also considering 100 repetitions). Sounds from fish used for the training set were also included in the testing set (known as speaker dependent tests in automatic speech recognition). Because there is individuality in boatwhistles (Amorim and Vasconcelos, 2008), we verified the ability of the system to recognize this sound type with 504 boatwhistles from four individuals not considered in the training data.

4. Evaluation of the recognition system

For each optimal alignment, the number of substitution errors (i.e., when one signal type is recognised as another signal type, S), and deletion errors (i.e., when a sound type occurs but is not detected by the system – a false negative, D) were determined. The performance of the recognition systems was then evaluated by computing the percentage of correctly recognized sounds (identification rate) using,

$$Identification\,rate = \frac{N - D - S}{N} \times 100\%,$$

where N is the total number of labels in the reference transcriptions. Notice that this measure ignores insertion errors. i.e., when a signal is detected by the system but it did not occur —a false positive (Young *et al.*, 2006). Boatwhistles from different fish occasionally overlapped as they often call in choruses. In such cases, the segmentation was not perfect resulting in apparent insertion errors that were in fact

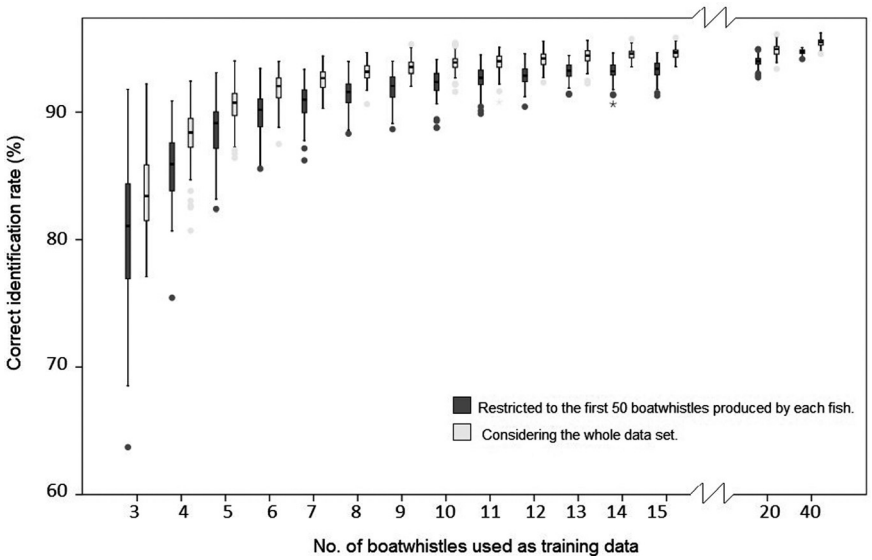

FIG. 3. Boxplots representing the relation between the identification rate of the system and the size of the training data set (from 3 to 40 boatwhistles). Each boxplot represent 100 repetitions. Dark gray symbols represent results using training data sets restricted to the first 50 boatwhistles produced by each fish. Lighter gray symbols represent the experiment with a training data set randomly selected considering the whole set of boatwhistles.

boatwhistles. As such, insertion errors were not considered in the evaluation of the recognition system (also see Parsons and Jones, 2000; Reby *et al*., 2006; Trifa *et al*., 2008).

III. RESULTS

A. Individual identification

Different frequency bandwidths and different signal processing feature sets (Table II) were used to classify the individual toadfish signals. All our individual identification trials presented high scores. Increasing the frequency bandwidth led to a classification improvement. For example, when considering 20–250 and 20–1000 Hz frequency bands, an identification rate of up to 88% and 92% was, respectively, obtained. The best recognition results were obtained with a 20–2000 Hz frequency range that led to ca. 95% correct classification.

Different signal processing feature sets held similar results. Using the feature sets including MFC with its delta and acceleration and energy coefficients resulted in a global identification rate of 95.0% ± 0.4% (mean ± standard deviation), the lowest from the three considered feature sets. Cepstrum instead of energy generated the highest identification rate achieving 95.5% ± 0.3% (Table III). With a feature set including PLP, energy and its delta and acceleration the system generated an identification rate of 95.2% ± 0.3%. The overall correct classification of each individual's boatwhistles (Table III) ranged between 65.2% and 100%, and 11 of 13 males presented rates above 90%, according to the confusion matrices obtained from 100 repetitions.

The effect of using different training sets is presented in Fig. 3. As expected, restricting the training to signals extracted from the first 50 boatwhistles produced lower mean identification rates than using the complete recordings.

TABLE III. Confusion matrix from the hidden Markov model classification computed using coefficients of MFC with its delta and acceleration, and cepstrum, considering a frequency range from 20 to 2000 Hz. The data set included 14 795 boatwhistles from 13 male Lusitanian toadfish produced during ca. 12 days. The model was trained with 40 randomly selected boatwhistles per male, and tested with the remaining ones. This procedure was repeated 100 times. The shown results are averages from these 100 repetitions. Please note that these averages are rounded and hence the grand total presented in the table is slightly different from the number of the boat whistles used; 95.5 ± 0.3% of tested boat whistles were correctly classified.

	Predicted group membership													
Fish	1	2	3	4	5	6	7	8	9	10	11	12	13	Percentage correct
1	398	0	0	0	0	0	0	0	0	0	0	0	0	100
2	14	101	4	1	1	3	0	0	23	0	0	1	7	65.2
3	0	1	1065	1	1	13	0	0	1	0	2	6	1	97.6
4	0	0	0	1514	4	0	0	0	0	136	1	8	2	90.9
5	0	0	5	6	1527	0	0	0	0	2	1	4	1	98.8
6	0	0	1	0	0	1542	0	0	0	0	1	10	0	99.2
7	18	0	2	2	9	33	704	2	3	1	36	1	1	86.7
8	0	0	0	0	3	0	20	609	0	0	3	1	0	95.8
9	26	0	0	0	0	10	2	0	475	0	1	0	3	91.9
10	1	0	8	86	5	1	8	2	0	1465	2	2	7	92.3
11	0	0	0	0	0	0	5	0	1	9	1567	0	0	99.1
12	0	0	0	0	7	0	0	0	0	0	0	1554	10	98.9
13	0	0	0	0	0	7	0	0	0	0	0	18	1576	98.4
Total														95.5 ± 0.3

In fact, training the model based on only seven boatwhistles randomly sampled within the whole data set resulted in at least 90% of correct identifications. However, if the training set was restricted to a random selection among the first 50 boatwhistles produced by each fish, more vocalizations were needed to achieve the same results (11 vocalizations). Increasing the number of vocalizations used for training the model resulted in asymptotic improvement of the mean identification rates along with decreased standard deviation.

B. Call type identification

The frequency bandwidth 20–2000 Hz and the feature set referred to in Table III were used for the call type identification system. This system achieved a higher identification rate when considering boatwhistles than for other sound types. A mean identification rate $98.4\% \pm 1.16\%$ was obtained in the identification of boatwhistles. In contrast, the other vocalizations were poorly recognized by the system with correct identification rates below 10%. Some mistakes in the classification of grunts were due to misidentifications associated with the last 100 ms of the boatwhistles. Using the 504 boatwhistles of other individuals not used for training the models, we obtained an identification rate of 99%. The remainder 1% was associated with some overlaps of boatwhistles produced simultaneously by different fish that were classified by the system as a single boatwhistles.

C. Duration of the boatwhistles

The duration of the boatwhistles estimated from our two recognition systems (individual identification and call type identification systems) was compared with the duration measured manually (Fig. 4). The duration estimated using the call type identification system was very similar to the values measured manually (paired Student's t-test; $n = 65$; $p > 0.05$), making this system a powerful tool to assess signal durations in large data sets. However, the individual identification system proved less adequate for the evaluation of the boatwhistles duration because the estimates were in this case significantly different from the manual measurements (paired Student's t-test; $n = 65$; $p < 0.001$).

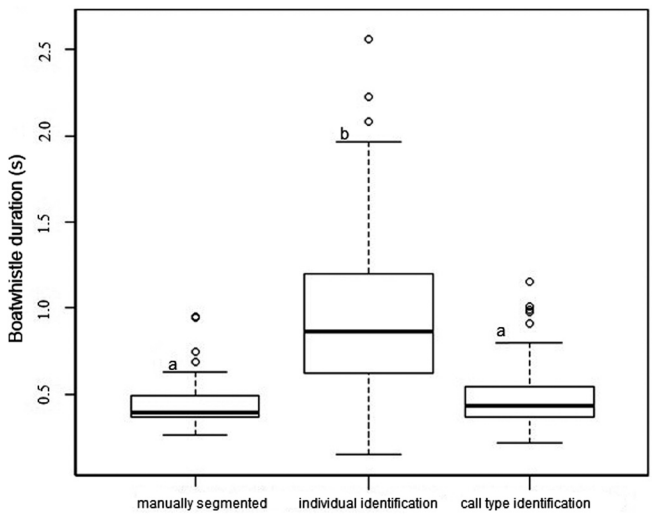

FIG. 4. Duration in seconds of 65 boatwhistles estimated by the user (manually segmented), by the individual identification system, and by the call type identification system computed on the MFC with cepstral, delta, and acceleration coefficients with a bandpass filter from 20 to 2000 Hz. Different letters represent pairwise significant differences ($p < 0.001$) using paired Student's t-tests.

IV. DISCUSSION

Our general goal was to develop a tool to study the vocalization activity of vocal fish in their natural habitat by using state of the art machine learning techniques for automatic individual identification and call recognition. For that, a highly vocal fish species with a large acoustic repertoire, the Lusitanian toadfish, was used as a model. This objective required the identification of several vocalization types and their assignment to individual males from round-the-clock sound recordings obtained during the Lusitanian toadfish breeding season, which lasts about 3 months. Recent major advances in automatic speech recognition have enabled the automatic analysis of bioacoustic signals (Table I), but to the best of our knowledge, the current work presents the first application of an automatic recognition HMM-based system to distinguish fish acoustic signals.

The results of HMM-based recognition systems using three feature sets to extract suitable features from boatwhistles of individual fish showed a good performance allowing a higher than 90% identification rate for the majority of toadfish males. Indeed the HMM system based on the feature set in Table III presented an average identification rate of 95.5%. This method reached values at least similar to or even higher than the identification rates observed in automatic recognition approaches with mammals' (Campbell *et al.*, 2000; Clemins, 2005; Reby *et al.*, 1997, 2006) and birds' (Terry and McGregor, 2002; Trawicki, 2005) vocalizations. The comparison, however, is not straightforward. These investigations were based on recordings involving a smaller number of animals singing simultaneously; this generally improves recognition (as exemplified in Trawicki, 2005). In addition, the acquisition of very good quality non-degraded sounds also improves recognition, such as the work carried out by Clemins *et al.* (2005), who recorded captive elephants with microphones held by collars.

One important issue of automatic sound recognition systems is the choice of data to train the models (Reby *et al.*, 1997; Tao *et al.*, 2008; Young *et al.*, 2006). As depicted in Fig. 3, a larger number of vocalizations used in the training phase usually improves the model's recognition ability. Thus a trade-off exists between the effort to manually extract the training set of vocalizations and the accuracy of the system. Our study also indicates the importance of using vocalizations extracted from longer recordings that better represent the existing data variability instead of using the first collected signals (see Amorim *et al.*, 2011).

Tests using several frequency bandwidths revealed that the essential signal information lies within 20 and 1000 Hz, a frequency range encompassing the spectral components in Lusitanian toadfish vocalizations (Amorim and Vasconcelos, 2008). Below 20 Hz, the recording was dominated by background noise. When considering a 20–250 Hz frequency

band to train the recognition system for individual discrimination, an identification rate of up to 88% was achieved. According to Vasconcelos *et al.* (Vasconcelos *et al.*, 2007; Vasconcelos *et al.*, 2011b), this frequency band corresponds to the range of best hearing sensitivity of Lusitanian toadfish. Moreover, auditory evoked potentials showed that amplitude modulation and spectral components of boatwhistles were well represented in the brain (Vasconcelos *et al.*, 2011a). Altogether, these results suggest that boatwhistles contain enough information to allow Lusitanian toadfish to distinguish conspecific individuals.

Our call type recognition system was effective to identify boatwhistles (with a mean identification rate of 98.4%), the most salient sound produced by male Lusitanian toadfish but was not able to correctly classify other sound types such as croaks, double-croaks, and grunts. Indeed the identification score of these sound types was too low (<10%) to consider automatic monitoring. Some heterogeneity in the recognition rates of different sound types of a species is commonly reported (Chesmore and Ohya, 2004; Jahns, 2008; Kogan and Margoliash, 1998; Parsons and Jones, 2000; Schön *et al.*, 2001; Trawicki, 2005). Several reasons may be responsible for the low identification rate of sounds (Young *et al.*, 2006). The low occurrence of croaks, double croaks, and grunts was the most likely factor because it did not allow proper adjustment of the model parameters. In contrast to boatwhistles, the pulsated nature of these sounds could have influenced misclassification. In fact, Young *et al.* (2006) point out that if parameters are adjusted to discriminate some sound characteristics, the system could overlook other sound types. Moreover, these sound types are also much shorter than boatwhistles precluding accurate HMM estimation. In contrast, long and stable vocalisations such as boatwhistles and hums found in other Batrachoidids (Bass and McKibben, 2003) are likely well recognised by this system.

A main parameter in the characterization of a species' acoustic signals is their duration. Our system for sound type recognition appeared to give an excellent estimation of the duration of boatwhistles. However, caution is recommended since errors can be expected caused by overlapping sounds from different individuals, misclassification of parts of boatwhistles, or unrecognized atypical boatwhistles due to much shorter or longer duration, noise interference, etc. These are challenging issues in automatic recognitions systems. For example, Stowell *et al.* (2013) addressed the recognition of overlapping calls in birds and Zhang and Gatica-Perez (2005), proposed a semi-supervised adapted HMM system using Bayesian statistics to minimise recognition error of infrequent sounds.

A recent study has applied automatic detection techniques to non-communicatory fish sounds by combining short-term spectral analysis with Gaussian mixture models (Diep *et al.*, 2013). Diep *et al.* (2013) aimed at counting shad spawning acts that are detectable by their associated splashing sounds. In contrast with fish acoustic signals used in social communication, these splashing sounds are non-stationary (i.e., the short-term spectral content is changing with time) and may notably differ from one to the other (Diep *et al.*, 2013), thus imposing very different solutions to the ones presented in the present study.

Here the usefulness of HMM-based automatic recognition systems to extensively analyse toadfish sound recordings is demonstrated. Future work using this system will allow assessing subtleties of the Lusitanian toadfish vocal behaviour in its natural habitat, which requires very long round-the-clock sound recordings, such as vocal rhythms, vocal interactions among fish, etc. Adjustments on this system to allow more subtle discriminations of sounds could, for example, permit to infer the dynamics of agonistic interactions throughout the breeding season by sorting advertisement from agonistic boatwhistles, which differ in dominant frequency and amplitude modulation (Vasconcelos *et al.*, 2010). Also the ability to recognize the less frequent sounds may improve by increasing data sets, which can be relevant to study this species communication.

The sounds used in the present study were registered in a complex natural estuarine environment not only presenting fluctuations of environmental parameters (e.g., temperature, turbidity, salinity, and light) but also affected by anthropogenic noise such as boat noise. Even so the recognition systems showed a high identification rate. However, in the present study, hydrophones were placed very close to sound producing animals. In a situation where the location of fish is not known, the identification rate could drop due to reduced SNR. However, Amorim and Vasconcelos (2008) used a discriminant function analysis and obtained a high classification rate of BWs from unseen individual Lusitanian toadfish recorded from a pier (i.e., with unknown distance from the hydrophone), suggesting that our automatic recognition system would perform well.

In summary, this recognition system could be an important tool in studies where analysis of very large recordings is required and invites future studies to expand it to other vocal fish species. For example, this system could be relevant to assess the distribution and abundance of commercial fish species such as the cod (Nordeide and Kjellsby, 1999) and the meagre (Lagardère and Mariani, 2006).

ACKNOWLEDGMENTS

This study was funded by the Science and Technology Foundation, Portugal (Project No. PTDC/MAR/118767/2010 and strategic Project No. UID/MAR/04292/2013 granted to MARE and UID/BIA/00329/2013 granted to cE3c). We would like to thank Daniel Alves and Carlotta Conti for their assistance during the field work allowing this investigation.

Acevedo, M., and Corrada-Bravo, C. (**2009**). "Automated classification of bird and amphibian calls using machine learning: A comparison of methods," Ecol. Inform. **4**, 206–214.

Amorim, M. C. P., Simões, J. M., Almada, V., and Fonseca, P. J. (**2011**). "Stereotypy and variation of the mating call in the Lusitanian toadfish, *Halobatrachus didactylus*," Behav. Ecol. Sociobiol. **65**(4), 707–716.

Amorim, M. C. P., Simões, J. M., and Fonseca, P. J. (**2008**). "Acoustic communication in the Lusitanian toadfish, *Halobatrachus didactylus*: Evidence for an unusual large vocal repertoire," J. Mar. Biol. Assoc. UK **88**, 1069–1073.

Amorim, M. C. P., and Vasconcelos, R. O. (**2008**). "Variability in the mating calls of the Lusitanian toadfish *Halobatrachus didactylus*: Cues for potential individual recognition," J. Fish Biol. **73**, 1267–1283.

Amorim, M. C. P., Vasconcelos, R. O., Marques, J. F., and Almada, F. (**2006**). "Seasonal variation of sound production in the Lusitanian toadfish *Halobatrachus didactylus*," J. Fish Biol. **69**, 1892–1899.
Anderson, S., Dave, A., and Margoliash, D. (**1996**). "Template based automatic recognition of birdsong syllables from continuous recordings," J. Acoust. Soc. Am. **100**(2), 1209–1219.
Baker, J. (**1975**). "The DRAGON system–An overview," IEEE Trans. Acoust. Speech Signal Process. **23**(1), 24–29.
Bass, A., and McKibben, J. (**2003**). "Neural mechanisms and behaviors for acoustic communication in teleost fish," Prog. Neurobiol. **69**, 1–26.
Baum, L., Petrie, T., Soules, G., and Weiss, N. (**1970**). "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," Ann. Math. Stat. **41**, 164–171.
Bridle, J., and Brown, M. (**1974**). "An experimental automatic word recognition system," Joint Speech Research Unit Report 1003.5, Malvern, UK.
Campbell, G., Gisiner, R., and David, A. (**2000**). "Acoustic identification of female Steller sea lions (*Eumetopias jubatus*)," J. Acoust. Soc. Am. **111**(6), 2920–2928.
Chesmore, D. (**2008**). "Automated bioacoustic identification of insects for phytosanitory and ecological applications," Comput. Bioacoust. Assess. Biodiversity **234**, 59–72.
Chesmore, E. D. (**2001**). "Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals," Appl. Acoust. **62**, 1359–1374.
Chesmore, E. D., and Ohya, E. (**2004**). "Automated identification of field-recorded songs of four British grasshoppers using bioacoustic signal recognition," Bull. Entomol. Res. **94**, 319–330.
Chou, C., Lee, C., and Ni, H. (**2007**). "Bird species recognition by comparing the HMMs of the syllables," in *Second International Conference on Innovative Computing, Information and Control*, Kumamoto, Japan, pp. 143–143.
Chou, C., Liu, P., and Cai, B. (**2008**). "On the studies of syllable segmentation and improving MFCCs for automatic birdsong recognition," in *IEEE Asia-Pacific Services Computing Conference*, Yilan, Taiwan, pp. 745–750.
Chu, W., and Blumstein, D. (**2011**). "Noise robust bird song detection using syllable pattern-based hidden Markov models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, pp. 345–348.
Clemins, P. (**2005**). "Automatic classification of animal vocalizations," Ph. D. dissertation, Marquette University, Milwaukee, WI, pp. 1–138.
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**(2), 956–963.
Dahl, G., Yu, D., Deng, L., and Acero, A. (**2012**). "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Trans. Audio Speech Lang. Process. **20**(1), 30–42.
Diep, D., Nonon, H., Marc, I., Delhom, J., and Roure, F. (**2013**). "Acoustic counting and monitoring of shad fish population," in *International AmiBio Workshop: Recent Progress in Computational Bioacoustics for Assessing Biodiversity*, Bonn, Germany, pp. 1–5.
Efron, B. (**1981**). "Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods," Biometrika **68**(3), 589–599.
Fonseca, P. J., and Maia Alves, J. (**2011**). Electret capsule hydrophone: A new underwater sound detector, Patent Application No. PT105, 933.
Forney, G. (**1973**). "The viterbi algorithm," Proc. IEEE **61**, 268–278.
Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. **87**(4), 1738–1752.
Jahns, G. (**2008**)."Call recognition to identify cow conditions—A call-recogniser translating calls to text," Comput. Electron. Agricul. **62**, 54–58.
Jelinek, F. (**1976**). "Continuous speech recognition by statistical methods," Proc. IEEE **64**, 532–556.
Jelinek, F., Bahl, L., and Mercer, R. (**1975**). "Design of a linguistic statistical decoder for the recognition of continuous speech," IEEE Trans. Inform. Theory **21**, 250–256.
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**(4), 2185–2196.
Küsel, E. T., Mellinger, D. K., Thomas, L., Marques, T. A., Moretti, D., and Ward, J. (**2011**). "Cetacean population density estimation from single fixed sensors using passive acoustics," J. Acoust. Soc. Am. **129**(6), 3610–3622.
Ladich, F. (**2004**). "Sound production and acoustic communication," in *The Senses of Fishes. Adaptations for the Reception of Natural Stimuli*, edited by G. van der Emde, J. Mogdans, and B. G. Kapoor (Narosa Publ. House, New Delhi), pp. 210–230.
Lagardère, J. P., and Mariani, A. (**2006**). "Spawning sounds in meagre *Argyrosomus regius* recorded in the Gironde estuary, France," J. Fish Biol. **69**, 1697–1708.
Lee, C., Lee, Y., and Huang, R. (**2006**). "Automatic recognition of bird songs using cepstral coefficients," J. Inform. Technol. Appl. **1**, 17–23.
Lippmann, R. (**1988**). "Neural network classifiers for speech recognition," Lincoln Lab. J. **1**(1), 107–124.
McDermott, E., Iwamida, H., Katagiri, S., and Tohkura, Y. (**1990**). "Shift-tolerant LVQ and hybrid LVQ-HMM for phoneme recognition," in *Readings in Speech Recognition*, edited by A. Waibel and K.-F. Lee (Morgan Kaufmann, San Francisco), pp. 425–438.
McIlraith, A., and Card, H. (**1995**). "Birdsong recognition with DSP and neural networks," in *Proceedings of WESCANEX95 Communications, Power, and Computing*, Winnipeg, Manitoba, Canada (IEEE Service Center, Piscataway, NJ), pp. 409–414.
Mellinger, D. K. (**2002**). ISHMAEL 1. 0 User' s Guide, NOAA, Technical Memorandum OAR PMEL-120, available from NOAA/PMEL/OERD, 2115 SE OSU Drive, Newport, OR 97365-5258, http://www.pmel.noaa.gov/pubs/PDF/mell2434/mell2434.pdf (Last viewed November 15, 2013).
Mills, H. (**1995**). "Automatic detection and classification of nocturnal migrant bird calls," J. Acoust. Soc. Am. **97**, 3370–3370.
Murray, S., Mercado, E., and Roitblat, H. (**1998**). "The neural network classification of false killer whale (*Pseudorca crassidens*) vocalizations," J. Acoust. Soc. Am. **104**(6), 3626–3633.
Nordeide, J. T., and Kjellsby, E. (**1999**). "Sound from spawning cod at their spawning grounds," ICES J. Mar. Sci. **56**, 326–332.
Oppenheim, A. V. (**1974**). *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ), pp. 1–784.
O'Shaughnessy, D. (**1987**). *Speech Communication: Human and Machine* (Addison-Wesley Series in Electrical Engineering, IEEE Press, Reading, MA), pp. 204–211.
Pace, F., White, P., and Adam, O. (**2012**). "Hidden Markov Modeling for humpback whale (*Megaptera novaeangliae*) call classification," Proc. Meet. Acoust. **17**(1), 070046.
Parsons, S. (**2001**). "Identification of New Zealand bats (*Chalinolobus tuberculatus* and *Mystacina tuberculata*) in flight from analysis of echolocation calls by artificial neural networks," J. Zool. **253**, 447–456.
Parsons, S., and Jones, G. (**2000**). "Acoustic identification of twelve species of echolocating bat by discriminant function analysis and artificial neural networks," J. Exp. Biol. **203**, 2641–2656.
Potamitis, I. (**2007**). "Automatic acoustic identification of crickets and cicadas," in *9th International Symposium on Signal Processing and Its Applications*, Sharjah, United Arab Emirates, pp. 1–4.
Potter, J., Mellinger, D., and Clark, C. (**1994**). "Marine mammal call discrimination using artificial neural networks," J. Acoust. Soc. Am. **96**(3), 1255–1262.
Pozzi, L., Gamba, M., and Giacoma, C. (**2010**). "The use of Artificial Neural Networks to classify primate vocalizations: A pilot study on black lemurs," Am. J. Primatol. **72**, 337–348.
Rabiner, L. R. (**1989**). "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE **77**, 257–286.
Reby, D., André-Obrecht, R., Galinier, A., Farinas, J., and Cargnelutti, B. (**2006**). "Cepstral coefficients and hidden Markov models reveal idiosyncratic voice characteristics in red deer (*Cervus elaphus*) stags," J. Acoust. Soc. Am. **120**, 4080–4089.
Reby, D., Lek, S., Dimopoulos, I., Joachim, J., Lauga, J., and Aulagnier, S. (**1997**). "Artificial neural networks as a classification method in the behavioural sciences," Behav. Process. **40**, 35–43.
Reynolds, D., and Rose, R. (**1995**). "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Proc. **3**, 72–83.
Schön, P., Puppe, B., and Manteuffel, G. (**2001**). "Linear prediction coding analysis and self-organizing feature map as tools to classify stress calls of domestic pigs (*Sus scrofa*)," J. Acoust. Soc. Am. **110**(3), 1425–1431.
Stowell, D., Saso, M., Bodana, J., and Plumbley, M. D. (**2013**). "Improved multiple birdsong tracking with distribution derivative method and markov renewal process clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings*, pp. 468–472.
Tao, J., Johnson, M. T., and Osiejuk, T. S. (**2008**). "Acoustic model adaptation for ortolan bunting (*Emberiza hortulana* L) song-type classification," J. Acoust. Soc. Am. **123**, 1582–1590.

Terry, A. M. R., and McGregor, P. K. (**2002**). "Census and monitoring based on individually identifiable vocalizations: The role of neural networks," Anim. Conserv. **5**, 103–111.
Trawicki, M. (**2005**). "Automatic song-type classification and speaker identification of Norwegian Ortolan Bunting (*Emberiza hortulana*) vocalizations," in *IEEE Workshop on Machine Learning for Signal Processing*, pp. 277–282.
Trifa, V. M., Kirschel, A. N., Taylor, C. E., and Vallejo, E. E. (**2008**). "Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models," J. Acoust. Soc. Am. **123**(4), 2424–2431.
van der Schaar, M., Delory, E., Català, A., and André, M. (**2007**). "Neural network-based sperm whale click classification," J. Mar. Biol. Assoc. UK **87**(1), 35–38.
Vasconcelos, R. O., Amorim, M. C. P., and Ladich, F. (**2007**). "Effects of ship noise on the detectability of communication signals in the Lusitanian toadfish," J. Exp. Biol. **210**, 2104–2112.
Vasconcelos, R. O., Carrico, R., Ramos, A., Modesto, T., Fonseca, P. J., and Amorim, M. C. P. (**2012**). "Vocal behavior predicts reproductive success in a teleost fish," Behav. Ecol. **23**, 375–383.
Vasconcelos, R. O., Fonseca, P. J., Amorim, M. C. P., and Ladich, F. (**2011a**). "Representation of complex vocalizations in the Lusitanian toadfish auditory system: Evidence of fine temporal, frequency and amplitude discrimination," Proc. R. Soc. Lond. B **278**, 826–834.
Vasconcelos, R. O., Simões, J. M., Almada, V. C., Fonseca, P. J., and Amorim, M. C. P. (**2010**). "Vocal behavior during territorial intrusions in the Lusitanian toadfish: Boat whistles also function as territorial 'keep-out' signals," Ethology **116**, 155–165.
Vasconcelos, R. O., Sisneros, J., Amorim, M. C. P., and Fonseca, P. J. (**2011b**). "Auditory saccular sensitivity of the vocal Lusitanian toadfish: Low frequency tuning allows acoustic communication throughout the year," J. Comp. Physiol. A **197**(9), 903–913.
Yen, G., and Fu, Q. (**2001**). "Automatic frog calls monitoring system: A machine learning approach," Int. J. Comput. Intell. Appl. **1**, 165–186.
Young, S., and Bloothooft, G. (**1997**). *Corpus-Based Methods in Language and Speech Processing* (Kluwer Academic, New York), pp. 1–235.
Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, P., Valtchev, V., and Woodland, P. (**2006**). *The HTK Book (for HTK Version 3.4)* (Cambridge University Press, Cambridge, UK), pp. 1–359.
Yu, H., and Oh, Y. (**1997**). "A neural network for 500 vocabulary word spotting using acoustic sub-word units," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, pp. 3277–3280.
Zhang, D., and Gatica-Perez, D. (**2005**). "Semi-supervised adapted HMMs for unusual event detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, pp. 611–618.