



An unsupervised Hidden Markov Model-based system for the detection and classification of blue whale vocalizations off Chile

Susannah J. Buchan, Rodrigo Mahú, Jorge Wuth, Naysa Balcazar-Cabrera, Laura Gutierrez, Sergio Neira & Néstor Becerra Yoma

To cite this article: Susannah J. Buchan, Rodrigo Mahú, Jorge Wuth, Naysa Balcazar-Cabrera, Laura Gutierrez, Sergio Neira & Néstor Becerra Yoma (2020) An unsupervised Hidden Markov Model-based system for the detection and classification of blue whale vocalizations off Chile, *Bioacoustics*, 29:2, 140-167, DOI: [10.1080/09524622.2018.1563758](https://doi.org/10.1080/09524622.2018.1563758)

To link to this article: <https://doi.org/10.1080/09524622.2018.1563758>



Published online: 15 Jan 2019.



Submit your article to this journal [↗](#)



Article views: 227



View related articles [↗](#)



View Crossmark data [↗](#)



An unsupervised Hidden Markov Model-based system for the detection and classification of blue whale vocalizations off Chile

Susannah J. Buchan^{a,b,c}, Rodrigo Mahú^d, Jorge Wuth^d, Naysa Balcazar-Cabrera^a, Laura Gutierrez^e, Sergio Neira^a and Néstor Becerra Yoma^d

^aCenter for Oceanographic Research COPAS Sur-Austral, University of Concepción, Concepción, Chile; ^bCentro de Estudios Avanzados en Zonas Áridas (CEAZA), Coquimbo, Chile; ^cWoods Hole Oceanographic Institution, Biology Department, Woods Hole, MA, USA; ^dSpeech and Processing Transmission Lab., Dept. of Electrical Engineering, Universidad de Chile, Santiago, Chile; ^eCentro de Investigación y Gestión de Recursos Naturales (CIGREN), Universidad de Valparaíso, Valparaíso, Chile

ABSTRACT

In this paper, we present an automatic method, without human supervision, for the detection and classification of blue whale vocalizations from passive acoustic monitoring (PAM) data using Hidden Markov Model technology implemented with a state-of-the-art machine learning platform, the Kaldi speech processing toolkit. 157.5 hours of PAM data were annotated for model training and testing, selected from a dataset collected from the Corcovado Gulf, Chilean Patagonia in 2016. The system obtained produced 85.3% accuracy for detection and classification of a range of different blue whale vocalizations. This system was then validated by comparing its unsupervised detection and classification results with the published results of southeast Pacific blue whale song phrase ('SEP2') via spectrogram cross-correlation, involving a dataset collected with a different hydrophone instrument. The proposed system led to a reduction in the root mean square error relative to published results as high as 80% when compared with comparable methods employed elsewhere. This is a significant step in advancing the monitoring of endangered whale populations in this region, which remains poorly covered in terms of PAM and general ocean observation. With further training, testing and validation, this system can be applied to other target signals and regions of the world ocean.

ARTICLE HISTORY

Received 17 July 2018
Accepted 12 December 2018


KEYWORDS

Blue whale vocalizations; unsupervised detection and classification; HMM; machine learning

1. Introduction

1.1. Conservation status

The conservation status of most baleen whale species remains of concern¹ after populations were decimated by historical commercial whaling (Rocha et al. 2014). Today, anthropogenic activities, such as collisions with ships (e.g. Laist et al. 2001; Vanderlaan and Taggart 2007; Neilson et al. 2012), underwater noise (e.g. Clark et al. 2009; Hatch et al. 2012; Rolland et al. 2012) and fishing (Trites et al. 1997; Read et al. 2006;

CONTACT Néstor Becerra Yoma  nbecerra@ing.uchile.cl

The underlying research materials for this article can be accessed at <http://www.lptv.cl/en/blue-whales-1/>.

Knowlton et al. 2012) continue to threaten populations via lethal and sub-lethal effects. Determining the distribution and seasonal movements of baleen whales is fundamental to understanding their ecology (e.g. Croll et al. 2005; Stafford et al. 2009; Buchan and Quiñones 2016), monitoring population trends (e.g. Branch et al. 2007; Bejder et al. 2016) and temporal and spatial distribution (e.g. Samaran et al. 2013; Davis et al. 2017), as well as overall ecosystem health (Moore 2008). This information is essential for developing conservation strategies and marine spatial and soundscape planning that ensure the continued protection of baleen whales (Redfern et al. 2013, 2017; Williams et al. 2015; Van Opzeeland and Boebel 2018).

1.2. Passive acoustic monitoring

Passive Acoustic Monitoring (PAM) is a useful and widely used method for monitoring the temporal and spatial presence of vocalizing whales and dolphins throughout the world's oceans (Mellinger et al. 2007; Van Parijs et al. 2009; Helble et al. 2015; Tripovich et al. 2015; Au and Lammers 2016; Nieukirk et al. 2016; Thomisch et al. 2016). In the case of baleen whales, their loud repetitive low-frequency vocalizations, often below 500 Hz, can be detected by hydrophones tens of kilometres from their source (Širović et al. 2007). Baleen whale vocalizations are distinct at species level and can also be distinct at sub-species and/or regional level (McDonald et al. 2006; Delarue et al. 2009).

Male blue whales (*Balaenoptera musculus*), in different regions, are known to produce one or more distinct stereotyped songs; monitoring these songs has revealed distinct spatial and temporal distributions of acoustic groups, although overlap does occur (McDonald et al. 2006; Stafford et al. 2011; Samaran et al. 2013; Buchan et al. 2014, 2015; Balcazar et al. 2017). Blue whale songs are made up of phrases, that are in turn made up of units (individual sounds), and it is the frequency (Hz) and duration (s) characteristics of song units and the pattern of song phrasing, e.g. A-B-A-B or A-B-C-A-B-C, (labelled ABC according to standard nomenclature in the literature) that distinguishes between regional song types (McDonald et al. 2006; Buchan et al. 2014). These songs can be heard throughout the migratory range of a population (Stafford et al. 1999a, 2001). Blue whale song is largely stable over time as shown for most blue whale song types (McDonald et al. 2006, 2009) including the Chilean or Southeast Pacific song type (comparing Stafford et al. 1999a with Buchan et al. 2014). There is however some intra-annual variation in song production (Oleson et al. 2007a), as well as a decrease over decadal timescales in the frequency of tonal song components (McDonald et al. 2009; Gavrilov et al. 2012). Male and female blue whales are also known to produce highly variable down-swept vocalizations known as 'D-calls', typically between approximately 40 and 75 Hz, that may be related to foraging (McDonald et al. 2001; Oleson et al. 2007a) and have so far not been found to possess regional differences.

Northern Chilean Patagonia is a known baleen whale feeding ground, primarily for blue whales (Hucke-Gaete et al. 2004; Buchan and Quiñones 2016; Galletti-Vernazzani et al. 2017), but also for humpback (*Megaptera novaeangliae*), sei whales (*Balaenoptera borealis*) and other cetaceans (Hucke-Gaete et al. 2010; Viddi et al. 2010). Previous passive acoustic studies have shown that the blue whales that feed in Chilean Patagonia have two unique song types known as Southeast Pacific 1 (SEP1) and Southeast Pacific 2 (SEP2), both described in detail in Section 2.2, and the latter being the dominant song type (Cumplings

and Thompson 1971; Buchan et al. 2014, 2015). Both are also heard in the eastern Tropical Pacific (Stafford et al. 1999a; Buchan et al. 2014, 2015). SEP2 songs are heard in this area between November and July, and singing peaks during April, in the austral autumn (Buchan et al. 2015). Antarctic (AA) blue whales songs (Buchan et al. 2018) and humpback whale songs (Español-Jiménez and van der Schaar 2018) have also been reported in this area, with singing reported mostly during the austral summer and autumn, respectively.

1.3. The importance of unsupervised detection and classification of whale vocalizations

The use of bottom-mounted hydrophones to monitor baleen whales allows year-round data collection without the economic and logistical constraints of boat-based data collection, which is particularly valuable in remote regions like Chilean Patagonia. PAM over years or decades generates large passive acoustic datasets that cannot be analysed manually in a timely manner but require automatic methods to detect and classify whale vocalizations (e.g. Mellinger et al. 2007). Ideally, these methods would be without human supervision to reduce to a minimum the amount of time a human analyst spends on the analysis. Also, the more variable the vocalization type, the more challenging it is to achieve a robust automatic detection method that neither misses (0 false negatives) nor confuses target signals (0 false positives). Noise in the marine environment from wind, marine traffic, seismic surveys, underwater earthquakes, and other sources (Hildebrand 2009) makes successful detection of vocalizations all the more difficult when target signal-to-noise ratio is low. At present almost all analytical methods for the detection and classification of whale vocalizations require some human supervision, and many methods require a significant amount of supervision, which is both time consuming and introduces human bias and error. This is the case, for example, of widely used spectrogram cross-correlation which measures the similarity between an input acoustic signal and a kernel of the target signal, both of which are represented as a spectrogram, and detection occurs when the time-frequency features of the input signal closely matches those of the template (Mellinger and Clark 2000). This is widely used for detecting stereotyped signals of baleen whales (Stafford et al. 1999a; Mellinger and Clark 2000; Samaran et al. 2013; Buchan et al. 2015) but does require significant analyst time to assess error (false and true negative and positive detections).

1.4. Kaldi speech recognition toolkit

Machine learning applied to the detection and classification of whale vocalizations in PAM data offers a promising solution to this problem that can reduce to zero the amount of time required by human supervision after the models have been trained using an annotated dataset (Brown and Smaragdis 2009; Dugan et al. 2010; Shamir et al. 2014). A Hidden Markov Model (HMM) is a machine learning technique that provides probabilistic models for sequences of data. HMM allows to model time varying processes as a sequence of states where each state represents stationary or quasi-stationary subprocesses. The variability of observed features within a given state is modelled with observation probabilities (see Section 2.3.2). HMM has been the most successful approach for solving very complex problems, such as speech recognition. HMM has

been previously used in bioacoustics (Ren et al. 2009; Ranjard et al. 2017), including the identification of bird species (Potamitis et al. 2014), and classification of mammal vocalizations (Agranat 2013; Scheifele et al. 2015; Putland et al. 2018). The Kaldi² speech recognition toolkit (Povey et al. 2011) provides a state-of-the-art platform to run HMM and deep learning experiments that has been widely employed by the speech recognition community worldwide, but to our knowledge has not been used for the detection and classification of whale vocalizations, or indeed any other non-human vocalization. This toolkit is open source and highly customizable, and allows replicable results. Kaldi runs on any Linux distribution, or on Cygwin or Mac OS X.

1.5. Objective

In this study, an automatic method for the detection and classification of blue whale vocalizations from passive acoustic data was developed with HMM technology using the Kaldi toolkit. Other low-frequency signals were also modelled and targeted, such as humpback whale vocalizations, ship noise, seismic events and platform noise due to mooring line strumming.

2. Materials and methodology

2.1. Study site and data collection

Two separate passive acoustic databases were used for a) model training/testing and b) detection and classification validation. The training/testing database was taken from passive acoustic data that were collected between January 2016 and February 2017 with a bottom-mounted SM3M Deepwater Song Meter hydrophone³ deployed in the Corcovado Gulf, Northern Chilean Patagonia (43°52S, 73°31W) at a depth of 170 m (with an acoustic release). Data collection was continuous in 30 min consecutive sound files, recorded at a sample rate of 4000 Hz. 157.5 hours (corresponding to 315 30-min sound files), were selected for annotation from this database. Selection and annotation of files is described in [Section 2.2](#). From here on, we will refer to this annotated data as ‘Corcovado-Songmeter’.

Validation experiments were done with a second dataset so that results could be compared with the published results in Buchan et al. (2015). In this case, continuous passive acoustic data collected in 2012 (from February to June) slightly north of the Song Meter deployment site but within the same general study area (43°31S, 74°26W) using a different instrument (Marine Autonomous Recording Unit, MARU⁴) recording at a different sample rate of 2000 Hz over a 5-month deployment due to battery constraints (Buchan et al. 2015). From here on, we will refer to this database as ‘Corcovado-MARU’.

2.2. Data annotation for training and testing

For model training/testing, the annotated ‘Corcovado-Songmeter’ database was used as input to train and test the HMM-based system. The files that were chosen were selected by an experienced bioacoustic analyst to obtain files with clear examples of noise (12% of all files) and whale vocalizations (88% of files). Files were selected from almost all months of the dataset, except for October and November because the analyst found no examples of whale

vocalizations or noise that were useful for training. The principal sources of noise identified were: diffuse background noise, ship noise, earthquakes, and mooring line strumming. In the selected files there were: 474 SEP1 phrases (adding up 4.5 hours of vocalizations); 4028 SEP2 phrases (39.7 hours); 192 AA phrases (0.6 hours); 2760 D-calls (2.3 hours); and 18 sequences of HB whale songs (0.9 hours considering inter-unit intervals). In numerous files, whale vocalizations were overlapped.

Annotation of sound files was carried out by two bioacoustic analysts as follows: Files were viewed as spectrograms in Raven Pro 1.5 (Bioacoustics Research Program 2012) using the following parameters: 8192 FFT, 80% overlap, Hann window, with a window set to view 100 Hz/120 s. Noise and whale vocalizations were marked with a box drawn around the target sound to include the entire sound in both frequency and time. A Raven 'Selection Table' was compiled to include the following data: 'Begin Time' (start time of signal), 'End Time' (end time of signal), 'Begin File' (the name of the file where the signal begins), 'End File' (the name of the file where the signal ends), 'Type' (see Table 1(a,b)), and 'Comments' (any other relevant observations).

SEP1 and SEP2 phrase and unit were identified based on song spectral descriptions by Buchan et al. (2014, 2015). SEP1 is a three-unit phrase (A-B-C) with an average total duration of approximately 34 s. Unit A has a mean peak frequency of 21 Hz and an average duration of 11.4 s; unit B, a mean peak frequency of 49 Hz and average duration of 9.2 s; and unit C, a mean peak frequency of 25 Hz and average duration of 9.5 s (Figure 1). SEP2 is a four-unit phrase (A-B-C-D) lasting on average 60s (Figure 2). Unit A has a mean peak frequency of 24 Hz and an average duration of 9.5 s; unit B, a mean peak frequency of 24 Hz and average duration of 13 s; unit C, a mean peak frequency of 26 Hz and average duration of 5 s; and unit D, a mean peak frequency of 24 Hz and average duration of 13 s (Figure 2). Exclusively for modelling purposes, each unit was annotated separately, except for SEP2 units B and C that were annotated together because there is no pause between them. AA phrases were identified as those described by Ljungblad et al. (1998) and Širović et al. (2004) and annotated as the complete Z-note with a duration of approximately 18 s and a mean peak frequency of 27 Hz (Figure 3). D-calls were identified as those described by McDonald et al. (2001) and Oleson et al. (2007a), ranging in frequency between 40 Hz and 75 Hz (Figure 4).

Other signals were also annotated: sequences of humpback whale song units, earthquakes, ship noise, mooring line strumming, and diffuse background noise. Humpback song sequences were identified based on visual comparison with published spectrograms from Chile (Español-Jiménez and van der Schaar 2018) and Brazil (Sousa-Lima et al. 2018), and via personal communications with Dr. Sousa-Lima. Identification of earthquakes, ship noise and strumming were based on published spectrograms, (Erbe et al. 2015; McKenna et al. 2012; Dziak et al. 2015, respectively). Annotation types are listed in Table 1(a,b), and examples of annotations can be seen in Figures 1–4.

Each type of annotated event (Table 1) was modelled with a three-state left-to-right without state skip transition HMM, Figure 5(a). In the case of humpback whale vocalizations, the vocalization type 'HB' in Table 1 denotes an entire sequence or cluster of humpback song units and was modelled as such with the HMM in Figure 5(a). This was because individual units of humpback whale song show a highly variable nature. In effect, humpback whale vocalizations vary between 20 Hz (Thompson et al. 1986) and 6 kHz (Stimpert et al. 2011), with harmonics

Table 1. Labels of events modelled with HMMs: (a) single whale vocalization types, (b) noise and other acoustic events, and (c) overlapping whale vocalizations.

(a) Single whale vocalization types.			
<i>Label</i>	<i>Description</i>		
'AA'	Antarctic blue whale song		
'D'	Blue whale D-call		
'S1.1'	Blue whale song SEP 1 unit A		
'S1.2'	Blue whale song SEP 1 unit B		
'S1.3'	Blue whale song SEP 1 unit C		
'S2.1'	Blue whale song SEP 2 unit A		
'S2.2'	Blue whale song SEP 2 units B and C		
'S2.3'	Blue whale song SEP 2 unit D		
'SEP'	Unidentified SEP blue whale song		
'HB'	Humpback whale song unit		

(b) Noise and other acoustic events.			
<i>Label</i>	<i>Description</i>		
'SIL'	Silence		
'UND'	Noise of undefined origin, including earthquakes		
'SHIP'	Ship noise		
'ST'	Strumming (platform noise)		

(c) Overlapping whale vocalizations.				
<i>Label</i>	<i>Description</i>		<i>Label</i>	<i>Description</i>
'AAD'	'AA' and 'D'		'S12S23'	'S1.2' and 'S2.3'
'AAS21'	'AA' and 'S2.1'		'S13S21'	'S1.3' and 'S2.1'
'AAS22'	'AA' and 'S2.2'		'S13S22'	'S1.3' and 'S2.2'
'AAS23'	'AA' and 'S2.3'		'S13S23'	'S1.3' and 'S2.3'
'AASEP'	'AA' and 'SEP'		'S13SEP'	'S1.3' and 'SEP'
'DS13'	'D' and 'S1.3'		'S21S22'	'S2.1' and 'S2.2'
'DS21'	'D' and 'S2.1'		'S21S23'	'S2.1' and 'S2.3'
'DS22'	'D' and 'S2.2'		'S21SEP'	'S2.1' and 'SEP'
'DS23'	'D' and 'S2.3'		'S21HB'	'S2.1' and 'HB'
'DSEP'	'D' and 'SEP'		'S22S23'	'S2.2' and 'S2.3'
'S11S12'	'S1.1' and 'S1.2'		'S22SEP'	'S2.2' and 'SEP'
'S11S22'	'S1.1' and 'S2.2'		'S22HB'	'S2.2' and 'HB'
'S11S23'	'S1.1' and 'S2.3'		'S23SEP'	'S2.3' and 'SEP'
'S12S13'	'S1.2' and 'S1.3'		'S23HB'	'S2.3' and 'HB'
'S12S22'	'S1.2' and 'S2.2'			

that can extend beyond 24 kHz (Au et al. 2006). In addition, song units appear in sequences that lack regularity. By modelling complete sequences, the HMM observation probability (Section 2.3.2) is expected to model the variability within the clusters. Blue whale D-calls also display frequency variability, but less so than for humpback vocalizations, i.e. between 40 Hz and 75 Hz (Oleson et al. 2007a). In addition, D-calls do not necessarily always appear in sequences and therefore were modelled individually with the three-state left-to-right without state skip transition HMM shown in Figure 5(a).

Two or more simultaneous whale vocalization types were observed in the 'Corcovado-Songmeter' database. Since the number of superimpositions of three or more units was very low (i.e. 0.14% of the total duration of data containing whale vocalizations in the 'Corcovado-Songmeter' database), we modelled the superimpositions of up to two song units (see Table 1(c)). The models of individual units (94.66%) and superimpositions of two units (5.20%) made up 99.86% of the total duration of data containing whale vocalizations

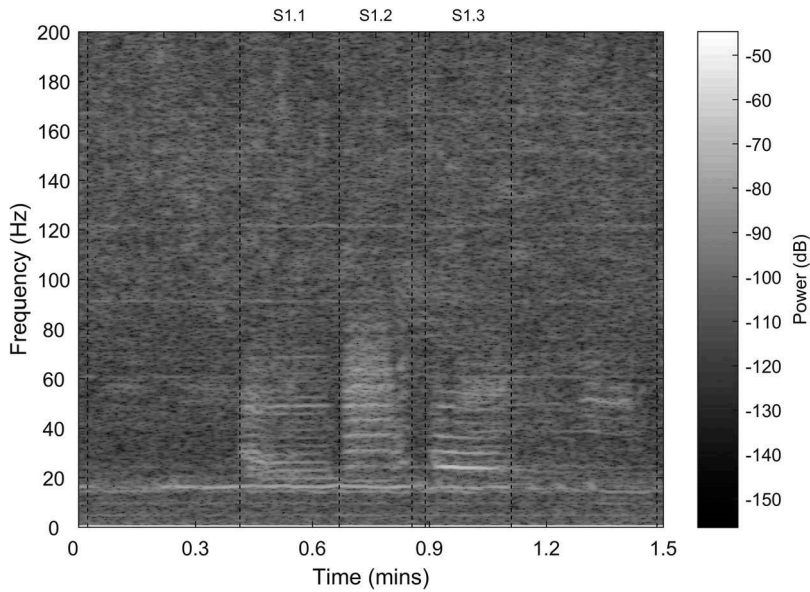


Figure 1. Example of annotation for SEP1.

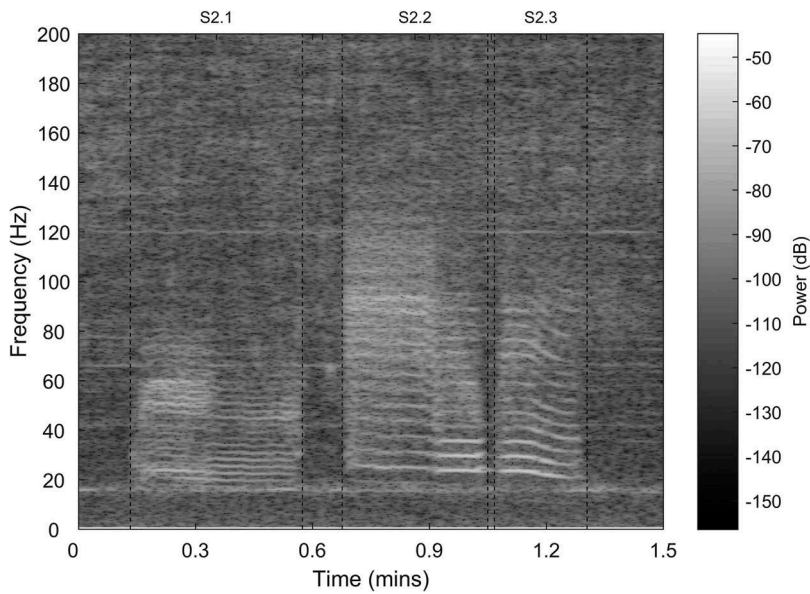


Figure 2. Example of annotation for SEP2.

present in the data. Modelling the remaining 0.14% of events would increase the number of vocalization types and the amount of training data would not be enough for their HMMs.

The HMM training procedure is composed of alignment-model estimation sequences. The alignment allocates a given interval of the signal to a model, and then this model is adapted according to the interval that was assigned to it. As a result, this corresponds to

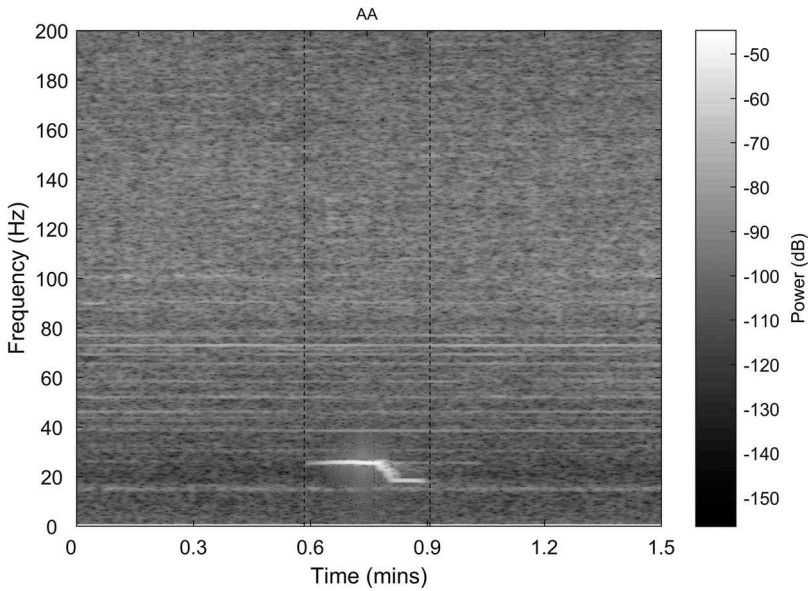


Figure 3. Example of annotation for Antarctic blue whale.

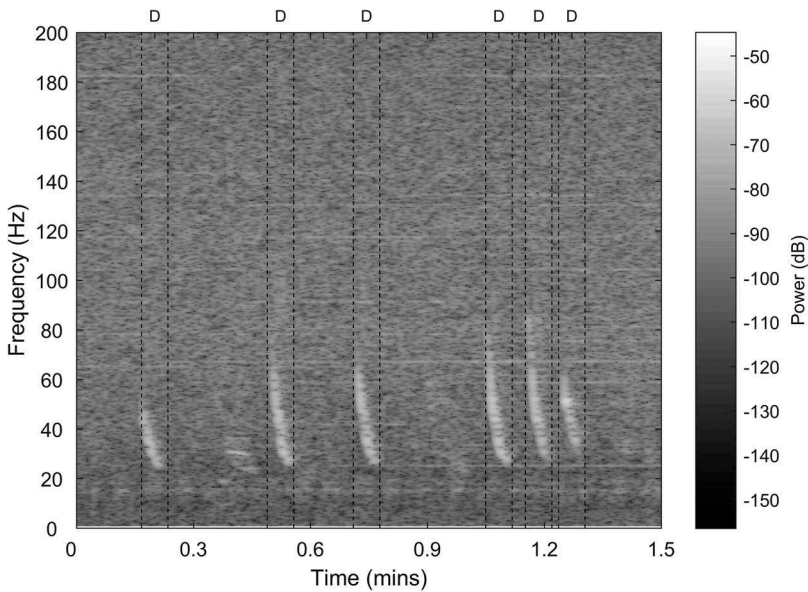


Figure 4. Example of annotation for blue whale D-calls.

a gradient based method and is highly dependent on the initial model in the first alignment of the training algorithm. In this paper, our strategy was to generate initial models trained with a set of annotated signals, making sure that these initial models would be as accurate as possible.

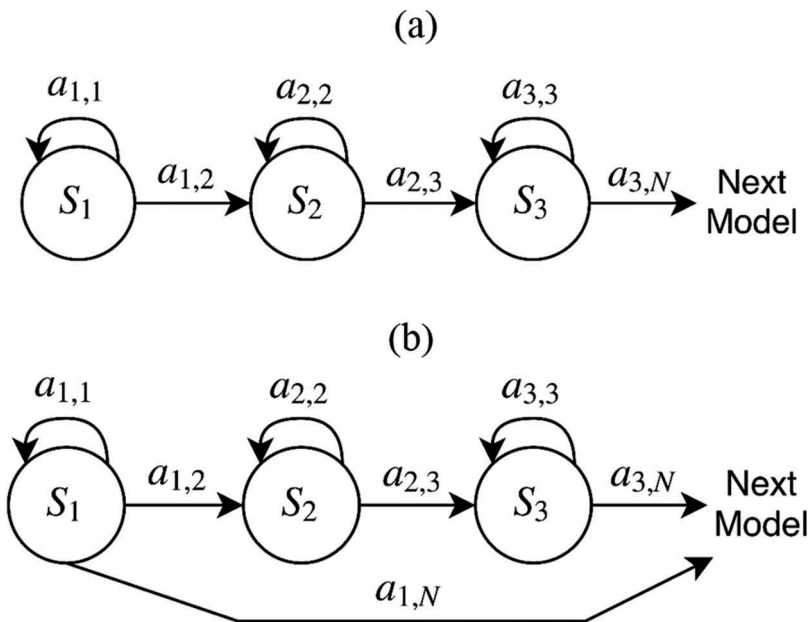


Figure 5. (a) Three-state left-to-right without state skip transition HMM topology employed to represent single and overlapping whale vocalizations according to Table 1(a,c). (b) Three state left-to-right with state skip transition HMM to model noise.

2.3. Introduction to Hidden Markov Models

An HMM is composed of a sequence of states that can model non-stationary signals as a sequence of pseudo stationary events. HMMs are a finite state machine defined by three sets of parameters: a) transition probabilities between states; b) observation probabilities in each state; and, c) initial probabilities for each state (Huang et al. 1990). In this study, whale vocalizations that were labelled according to Table 1(a,c) were modelled using the three-state left-to-right without state skip transition HMM topology (Figure 5(a)) to represent their dynamics. The first, second and third states modelled the beginning, middle and end of a whale vocalization, respectively. As explained in more detail later, the recorded signal was divided into short-term windows, and within each window a set of features was estimated (i.e. feature extraction). The set of features estimated for each window is denominated frame. Consequently, each frame t is represented by a feature vector denoted as \mathbf{O}_t that is allocated to one of the corresponding states.

2.3.1. Transition probabilities

In an HMM, the current state may change from one frame to the following one, and the probability for changing to state j from state i is given by the transition probability a_{ij} . Allowed state transitions were drawn as arcs in the model (Figure 5(a,b)). For instance, in the HMM of Figure 5(a), given a state i , only state transitions to the same state i or state $i + 1$ were possible.

2.3.2. Observation probabilities

Given an HMM λ and a feature vector \mathbf{O}_t , to each state S_i corresponds an observation probability, $\Pr(\mathbf{O}_t|S_i, \lambda)$. Here, the observation probability was modelled with a probability density function represented by a Gaussian Mixture Model (GMM) composed of G Gaussians. The observation probability can be defined as (Huang et al. 1990):

$$\Pr(\mathbf{O}_t|S_i, \lambda) = \sum_{g=1}^G \phi_{g,i,\lambda} \times \mathfrak{N}(\mathbf{O}_t, \boldsymbol{\mu}_{g,i,\lambda}, \Sigma_{g,i,\lambda}) \quad (1)$$

where λ denotes a given HMM corresponding to a specific whale vocalization or, as explained later, to a type of event (see Table 1); G is the number of Gaussians per state; $\mathfrak{N}(\cdot; \boldsymbol{\mu}, \Sigma)$ is a multivariate Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ :

$$\mathfrak{N}(\mathbf{O}_t; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(\mathbf{O}_t - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{O}_t - \boldsymbol{\mu})} \quad (2)$$

where n is the dimensionality of vector \mathbf{O}_t , and $\phi_{g,i,\lambda}$ are the weights of the Gaussians.

2.3.3. Initial probabilities

Another parameter that defines an HMM is the vector of initial probabilities for each state $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_{N_S})$, where N_S is the number of states within the HMM (Rabiner and Juang 1986). Vector $\boldsymbol{\pi}$ represents the probability distribution of the initial state, i.e. the probability of a given state being assigned to the first frame of an event (i.e. whale vocalization or noise). Here, the first frame within an event was allocated to the first state in the corresponding HMM, i.e. $\boldsymbol{\pi} = (1, 0, 0)$.

2.3.4. Feature extraction

The goal of the feature extraction process is to reduce the data dimensionality by converting the sampled waveform into a sequence of parameter vectors with less redundant information. The feature extraction process is carried out by arranging the signal into frames, usually overlapping, by employing a Hamming window. For each frame, different features can be extracted to compose the observation vector of each frame. Accordingly, a recorded continuous signal, whose length is equal to T frames, is represented by a sequence of observation vectors $\mathbf{O}_1, \mathbf{O}_2, \dots, \mathbf{O}_t \dots \mathbf{O}_T$, where each vector is composed of features that characterize the corresponding process. Examples of popular parameters are cepstral (Huang et al. 1990) and linear prediction filter (LPC) (Esposito et al. 2013) coefficients. These kinds of features are called static because they represent information from a given frame.

Delta (Δ) and delta-delta ($\Delta\Delta$), known as differential and acceleration coefficients, are usually included to account for the temporal evolution of the static features described above. This time evolution is established on the basis of the first and second derivatives of the corresponding static features. Δ and $\Delta\Delta$ for the static features in frame t can be as defined as follows (Gold et al. 2011):

$$\Delta(t) = \frac{2 * \text{static_features}(t+2) + \text{static_features}(t+1) - \text{static_features}(t-1) - 2 * \text{static_features}(t-2)}{10} \quad (3)$$

$$\Delta(t) = \frac{2\Delta(t+2) + \Delta(t+1) - \Delta(t-1) - 2\Delta(t-2)}{10} \quad (4)$$

where $static_features(t)$ denotes the corresponding static feature vector in frame t .

Mean and variance normalization (MVN) (Li et al. 2015) of the coefficients can also be employed. MVN reduces the distortion brought by additive noise and convolutive channel, and it is applied to all the feature vectors along each event signal.

2.3.5. Training and decoding

Following feature extraction, the HMM parameters for each whale vocalization and noise event were estimated from a set of recorded signals, i.e. the training dataset. Trained HMMs were employed in a decoding algorithm, called the Viterbi algorithm (Viterbi 1967; Rabiner 1989; Huang et al. 1990; Rabiner and Juang 1993), that processes a given testing signal to both detect and classify the type of events modelled here (Table 1). Consequently, given an input audio signal, the Viterbi algorithm allocates each frame to one of the HMMs and states that makes up the entire HMM network (see Section 3). This network was defined by the interconnection of all the HMMs that represent the whale vocalizations and noise event types in Table 1. In other words, each sequence of whale vocalizations and noise observed in the database can be generated by the HMM network in Figure 7. The training and testing datasets were generated by dividing the annotated ‘Corcovado-Songmeter’ database into two disjoint subsets, where the training dataset contained all the modelled event types (Table 1). The decoding procedure is both a detection and a classification procedure because it delivers the most likely sequence of whale vocalizations and noise in an input signal.

Figure 6 shows the system architecture employed here. Representative features from the whole dataset were extracted and the HMMs were trained: the transition probabilities and the Gaussian distributions for the observation probabilities were estimated from the training data by using an iterative procedure. The annotated training database was employed to define an accurate starting condition or initial models for the GMM training. In the decoding process, for a test signal, the optimal alignment was determined with the Viterbi algorithm. The optimal alignment is the most likely sequence of HMMs and states that can be assigned to the sequence of frames.

3. HMM network

All whale vocalization and noise event types in the annotated data were modelled in this study. The absence of whale vocalizations was represented by background noise. In this paper, the background noise was modelled by making use of a three-state left-to-right HMM with state skip transition according to Figure 5(b). This model allows the separation between contiguous whale vocalization units to be as short as one frame. Additionally, in order to examine the benefits of the proposed noise model, experiments were also carried out using the three-state left-to-right without state skip transition HMM topology shown in Figure 5(a).

Each acoustic event type (i.e. whale vocalization and noise) was associated with an HMM, and all the defined HMMs made up a network that represents all possible sequences of acoustic events, as shown in Figure 7. According to the proposed HMM

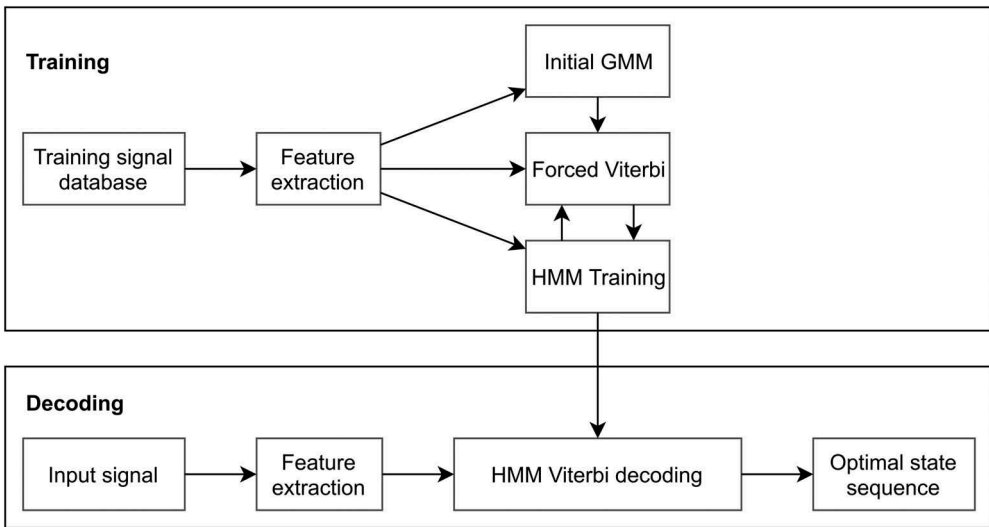


Figure 6. System architecture employed in this research.

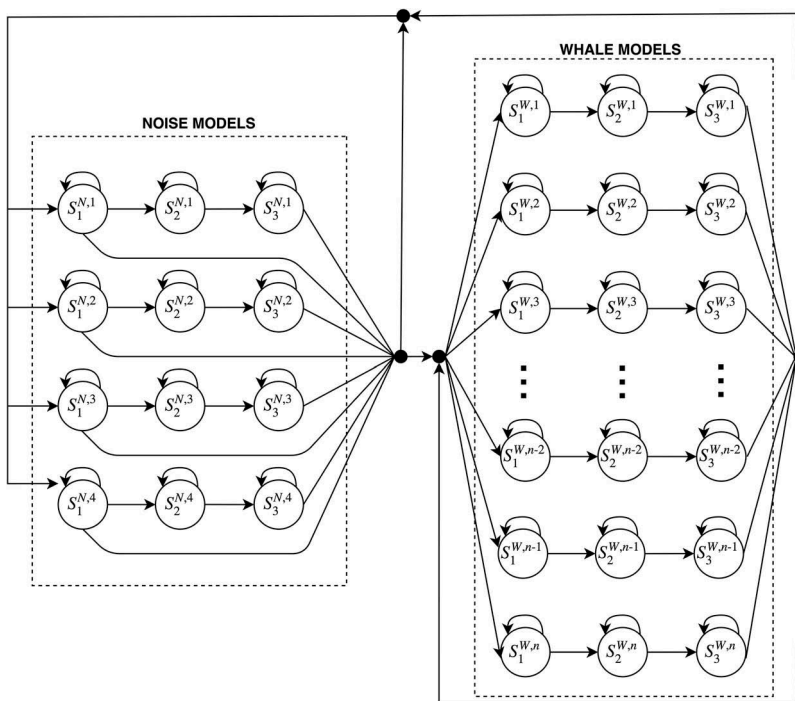


Figure 7. HMM network for the detection and classification of acoustic events. The network in this figure considers that noise is modelled with the three state left-to-right HMM with state skip transition. Where: $S_i^{W,\lambda}$ and $S_i^{N,\lambda}$ denote states within whale and noise models, respectively; and, λ and i correspond to the model and state indexes, respectively.

network, any whale vocalization can be preceded and followed by either another whale vocalization or a noise event. Two consecutive whale vocalizations can be observed without noise frames between them. The transition from one state to the next one is defined in such a way that after the first state of a whale vocalization or noise model, the following frames are allocated to the first or second state of the corresponding HMM according to the left-to-right without state skip transition topology (Figure 5(a)). Similarly, the transition from the third state within an HMM can be to the same state or towards the first state of any of the models that compose the network. By using the state skip transition topology (Figure 5(b)) to model noise, an additional transition is allowed from the first state of a noise model to the first state of any of the models that compose the network. For instance, after a noise event, a single or overlapped whale vocalization (Table 1(a,c)) can be detected. After that, a new noise event or whale vocalization can take place again.

4. Experimental setup and system description

Initial training/testing experiments were carried out with the ‘Corcovado-Songmeter’ database to evaluate different detection and classification platform configurations in terms of accuracy and sensitivity. They were performed with two Subsets (1 and 2) of the database. First, experiments were carried out with Subset 1 that contained 157 sound files for training and included all the models present in the annotated database (see Table 1), and the remaining 158 sound files (Subset 2) were used for testing. Then, experiments were performed with Subset 2 for training and Subset 1 for testing. In the latter case, five sound files were moved from Subset 1 to Subset 2 because they contained the only examples for ‘AAS11’, ‘AASEP’, ‘S12SEP’, ‘S21HB’, and ‘S22HB’ event types.

Once the detection and classification platforms were obtained, validation experiments were done by: 1) training the resulting platforms with the entire annotated ‘Corcovado-Songmeter’ database; 2) testing with the ‘Corcovado-MARU’ database; and 3) comparing the published results by Buchan et al. (2015) of the automatic detection of SEP2 phrases in the ‘Corcovado-MARU’ database using spectrogram cross-correlation. Buchan et al. (2015) analyzed acoustic data as follows: spectrograms were made using XBAT (Extensible Bioacoustic Tool; Bioacoustics Research Program 2012) with FFT: 4096 samples, 25% overlap, Hann window. Automatic detection in XBAT was carried out via spectrogram correlation, which quantifies the similarity between a signal and a template or kernel of a target sound (Mellinger and Clark 2000). The kernel used was units C and D from an SEP2 exemplar taken from the dataset. To assess true positives and true negatives, where the number of detections per month was fewer than 500, each detection was scanned visually and deleted if a false positive (incorrect detection). Otherwise, the first 48 h of data of every month were scanned visually to determine the number of false positives and false negatives (missed target sounds) as a percentage of the total number of detections. 3% of detections by the SEP2 detector were false positives and 28% of the SEP2 detections were false negatives. The number of corrected detections was calculated by subtracting the percent of false positive detections from initial detections (except for those months with 500 detections or fewer where all detections were reviewed) and adding the percent of false negatives on to all months.

Here, published results by Buchan et al. (2015) as average number of corrected detections per day of monitoring effort for each month of data were used as a ground truth of SEP2 detection and classification for platform validation.

4.1. Feature extraction and a new frequency compression curve

A feature extraction procedure was performed as mentioned in Section 2.3.4. For model training/testing experiments, each signal of the ‘Corcovado-Songmeter’ database was divided into 8192 sample frames with 50% overlap using a Hamming window. For each frame, the Fourier power spectrum was obtained by computing an 8192-point discrete Fourier transform. Given the sample rate of 4000 Hz, the frequency resolution of the discrete Fourier transform was equal to 0.4883 Hz. Because the blue whale vocalizations targeted in this study were below 200 Hz, we studied the effect of truncating the high frequency portion of the spectrum. The power spectrum was truncated to its first N samples, limiting the maximum frequency observable in the truncated spectrum. The optimal number of spectrum samples to be considered was determined empirically. Features were extracted from the truncated spectrum by employing a filterbank composed of different triangular filters similar to the one shown in Figure 8. Each filter gain F_i had central frequency fc_i and bandwidth B_i , and can be expressed as:

$$F_i = \begin{cases} -\frac{2}{B_i}|f - f_i^c| + 1 & \text{when } |f - f_i^c| \leq \frac{B_i}{2} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

The triangular filters were arranged in the target frequency range so that the lowest frequency f_i^{lowest} at each filter gain F_i corresponded to the central frequency fc_{i-1} of the previous filter gain F_{i-1} . Consequently, the filterbank shown in Figure 8 was obtained. The number of filters that composes a filterbank depends on the bandwidth of each triangular filter and on the bandwidth of interest in the truncated spectrum.

Here, the bandwidth of the filters was chosen to provide higher resolution at low frequencies than at higher frequencies. This is the rationale behind the Mel scale (Stevens and Volkman 1940), based on psychoacoustic experiments that are widely employed in speech processing. In effect, the Mel scale has been previously employed in bioacoustics, including bird vocalization analyses (Ranjard et al. 2015), classification of anurans (Noda et al. 2016), individual identification of Bornean male orangutans (Spillmann et al. 2017), and detection of whale vocalizations (Putland et al. 2018).

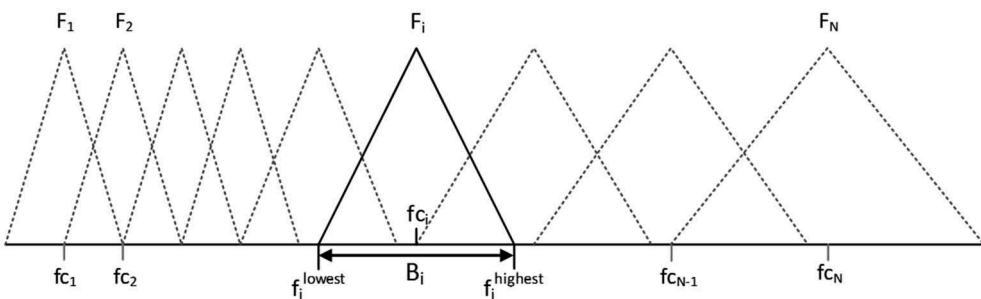


Figure 8. Filterbank composed of N triangular filters.

However, there is no reason to believe that the processing of whale vocalizations and human speech should share the same optimum frequency compression curve. In this paper the bandwidth for all the filters with fc_{i-1} below a threshold frequency f_{th} was kept constant and equal to B , while the bandwidth for filters with fc_{i-1} above f_{th} increased linearly according to fc_{i-1} . For the first filter, bandwidth B_1 and central frequency fc_1 are given by:

$$B_1 = B \quad (6)$$

$$fc_1 = \frac{B}{2} \quad (7)$$

For $i > 1$, given that the lowest frequency of filter F_i corresponds to the central frequency of filter F_{i-1} , B_i and fc_i can be described recursively as follows:

$$B_i = \begin{cases} B & \text{if } fc_{i-1} \leq f_{th} \\ B + (fc_{i-1} - f_{th}) \tan \alpha & \text{otherwise} \end{cases} \quad (8)$$

$$fc_i = fc_{i-1} + \frac{B_i}{2} \quad (9)$$

where f_{th} is the frequency over which the bandwidth begins to increase linearly with fc_i ; and, α is the constant that controls the frequency compression rate above fc_i . The bandwidth B was made equal to 0.98 Hz, which corresponds to two FFT samples. The proposed parametrization had three parameters that need to be tuned: N , the number of points (or bandwidth) of the truncated power spectrum; f_{th} , the frequency over which the bandwidth begins to increase linearly with respect to the filter central frequency; and α , the constant that controls the frequency compression curve above f_{th} . The optimal values for N , f_{th} and α were tuned with the ‘Corcovado-Songmeter’ database.

By applying the filterbank to the truncated power spectrogram, a set of static features was obtained corresponding to the log-energy at each filter by making use of the frequency compression curve described by Equations (8) and (9). Finally, the first (Δ) and second ($\Delta\Delta$) derivatives of the static features, as defined in Equations (3) and (4), were also estimated. MVN was applied to all the feature vectors along each recording. This parametrization will be referred to as ‘Non-linear-param’.

To compare the effect of the frequency compression curve, a special case of the previous parametrization was considered with $\alpha = 0$, i.e. suppressing the frequency compression curve and making all the filter bandwidths equal to B . This parametrization will be referred to as ‘Linear-param’ and depended exclusively on N . Additionally, a third set of features based on the Mel scale was estimated as in Putland et al. (2018): 24 static cepstral parameters plus Δ and $\Delta\Delta$ features per frame provided a final feature vector of 72 coefficients. MVN was applied to all the feature vectors along each recording. This parametrization will be referred to as ‘Mel-cepstral’.

For validation, models obtained with the ‘Corcovado-Songmeter’ database were applied to analyse the ‘Corcovado-MARU’ database. However, these databases were not recorded at the same sample rate (4000 Hz for the ‘Corcovado-Songmeter’ and 2000 Hz for the ‘Corcovado-MARU’). In order to keep the same frequency resolution of the discrete Fourier transform in both databases parametrization, each signal of the

‘Corcovado-MARU’ database was divided into 4096 sample frames with 50% overlap using a Hamming window. For each frame, the Fourier power spectrum was obtained by computing a 4096-point discrete Fourier transform. In this way, the frequency resolution of the discrete Fourier transform was equal to 0.4883 Hz, the same resolution employed in the ‘Corcovado-Songmeter’ database. Additionally, both databases were recorded with different hydrophones. This incorporates a channel mismatch that is reduced by applying MVN.

4.2. Training procedure and the initial models

As mentioned above, all the signals from the ‘Corcovado-Songmeter’ database were annotated by analysts. For training, all whale vocalizations and noise events were uniformly divided into three non-overlapping segments of feature vectors that were allocated to states 1, 2, and 3 sequentially within their corresponding HMM. The Gaussian distributions for the observation probabilities for each model were initially estimated by using the frames allocated in each state.

The initial transition probabilities for the whale models were $a_{11} = 0.75$, $a_{12} = 0.25$, $a_{22} = 0.75$, $a_{23} = 0.25$, $a_{33} = 0.75$ and $a_{3N} = 0.25$, and the transition probabilities for the noise models were $a_{11} = 0.45$, $a_{12} = 0.45$, $a_{1N} = 0.10$, $a_{22} = 0.5$, $a_{23} = 0.5$, $a_{33} = 0.75$ and $a_{3N} = 0.25$ (a_{1N} and a_{3N} denote the transition to another HMM in Figure 7). The forced Viterbi algorithm was used to assign each frame to a state in a given sequence of models. After this assignment, the HMMs parameters were updated with the frames in each state. The forced Viterbi algorithm and HMM update was repeated 40 times by default. At the start of the training procedure in Kaldi, all observation probabilities were composed of a single Gaussian. As the training procedure iterated, the number of Gaussians in each state could increase with respect to the number of frames allocated to the state, then the frames were reassigned within the GMM and the Gaussian parameters were re-estimated. In the default Kaldi GMM initialization procedure, Gaussians are initialized by uniformly distributing the frames to each model and state present in the audio transcription (i.e. the sequence of vocalizations and noise in the signal). As proposed here, a better initial frame-to-state assignment can be obtained by using the start and end time of each event type from the labelled training database. The two model initializations, i.e. the standard Kaldi GMM initialization (Semi-supervised-initialization) and the one proposed here (Fully-supervised-initialization) were compared.

4.3. Performance metric

In the testing procedure, the optimal alignment was estimated by applying the decoding Viterbi algorithm to each testing signal. The performance of the proposed HMM-based whale vocalization detection and classification system in the ‘Corcovado-Songmeter’ database was evaluated using frame level classification accuracy and sensitivity. The classification accuracy is the ratio between the number of frames with correct detection and classification of labelled events, and the total number of frames in the testing data. Consequently, error rate is 1 minus classification accuracy. Sensitivity is defined as $\frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})}$ (Putland et al. 2018) and was estimated with respect to SEP2 for

compatibility with the validation results on the ‘Corcovado-MARU’ database. In our HMM-based decoding system, according to Table 1, a true positive event was defined as the coincidence of a detected ‘S2.3’ event with one of the annotated reference labels ‘S2.1’, ‘S2.2’ or ‘S2.3’. For comparison reasons, classification accuracy or relative error rate difference were adopted because they provide a more complete description of the system performance and are widely employed in the literature. The validation experiments were carried out on the ‘Corcovado-MARU’ database. The average daily number of SEP2 detections per month were compared among the three different system configurations and the ‘reference’, i.e. the results published in Buchan et al. (2015). The system performance was evaluated in terms of the root mean square error (RMSE) between the bars of the reference results and each one of the evaluated configurations. The number of SEP2 phrase detections for each system configuration evaluated here was computed by counting the detected S2.3 events (Table 1(a)).

5. Results and discussion

From training and testing experiments with the ‘Corcovado-Songmeter’, the proposed fully-supervised HMM initialization with the non-linear parametrization in combination with the three-state left-to-right with state skip transition HMM to model noise achieved the highest classification accuracy and sensitivity (85.3% and 79.8%, respectively). The fully-supervised HMM initialization alone was able to lead to a reduction in error rate equal to 58.9% relative when compared to the ordinary semi-supervised HMM initialization provided by Kaldi. The optimal non-linear parametrization provided an error reduction of 8.7% relative when compared to the optimal linear-param. It is worth emphasizing that Mel-cepstral, which has been employed in several bioacoustic tasks, led to an increase in classification error of 169% compared to the optimal non-linear parametrization. In terms of the noise models, the three-state left-to-right with state skip transition HMM provided a reduction in error rate equal to 3.3% relative when compared to the topology without state skip transition. Finally, in the validation experiments with the ‘Corcovado-MARU’ database, the proposed system led a RMSE that is at least 76% lower than all the systems evaluated here. We discuss these results below.

5.1. Parametrization evaluation

The three parametrizations were evaluated on Subsets 1 and 2 from the ‘Corcovado-Songmeter’ database (see Section 4.1). The training method employed here is the proposed Fully-supervised-initialization. First, the optimal set of parameters of the Non-linear-param (i.e. N , f_{th} and α) was determined by a grid search with: $N = [300, 400, 500]$, which corresponds to 146 Hz, 195 Hz and 244 Hz, respectively; $f_{th} = [10 \text{ Hz}, 20 \text{ Hz}, 30 \text{ Hz}, 40 \text{ Hz}, 60 \text{ Hz}, 70 \text{ Hz}, 80 \text{ Hz}]$; and, $\alpha = [15, 30, 45, 60]$. The highest average classification accuracy was equal to 85.3% when $N = 300$ samples, $\alpha = 30$, and $f_{th} = 80$ Hz. The average classification accuracies vs. f_{th} and α with $N = 300$ are shown in Figure 9. These results suggest that a higher number of FFT samples does not lead to an increase in detection and classification accuracy. This may be due to the fact that the high frequency portion of the spectrum does not provides useful information to discriminate between the SEP1, SEP2 and

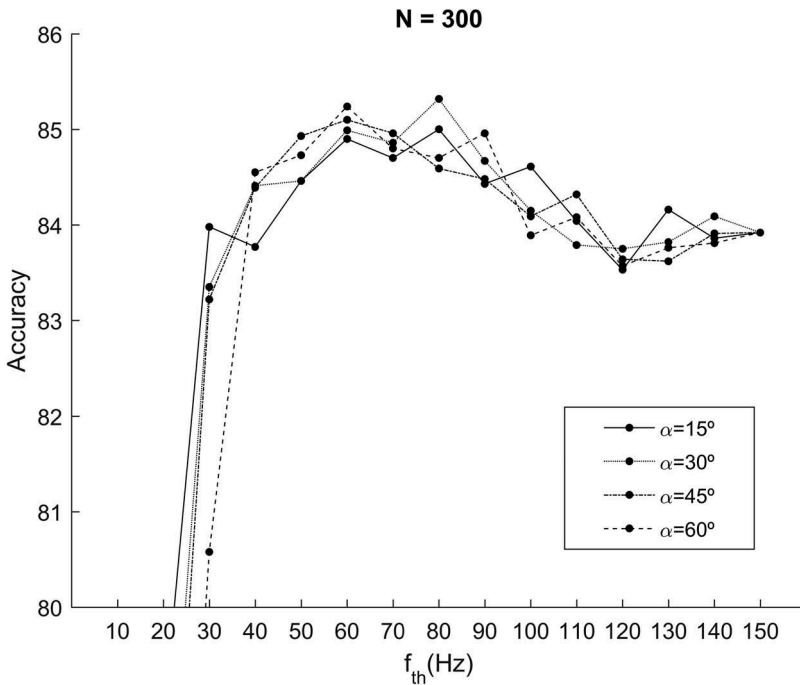


Figure 9. Classification accuracies obtained with non-linear-param and N made equal to 300. The HMM training procedure made use of fully-supervised-initialization. The classification accuracy decreases dramatically when $f_{th} < 30$ Hz.

AA blue whale song type. **Figure 9** shows that a higher detection and classification accuracy is achieved when the frequency compression curve starts at 40 Hz or 50 Hz, i.e. $f_{th} = 40$ Hz or $f_{th} = 50$ Hz. The optimum observed f_{th} was around 80 Hz. This result indicates that at lower frequencies (i.e. < 50 Hz), a higher frequency resolution leads to higher accuracy. In fact, classification accuracies decreased dramatically with $f_{th} < 30$ Hz. This is probably because these vocalizations are highly stereotyped and regular at low frequencies. **Figure 9** also shows that if frequency compression is delayed, i.e. $f_{th} > 80$ Hz, classification accuracies also decreased. It is worth highlighting that with higher accuracy comes greater discrimination among SEP1, SEP2 and AA.

Experiments were carried out to compare the performance of the tuned Non-linear-param, with the Linear-param and Mel-cepstral parametrizations. **Table 2** shows the average classification accuracies on ‘Corcovado-Songmeter’ database to compare the proposed Non-linear-param with Linear-param (i.e. $\alpha = 0$ in Equation (8) where all the filters in **Figure 8** have the same bandwidth) and Mel-cepstral. As shown in **Table 2**, the classification accuracy was reduced when N increased, and the highest accuracies, 85.3% and 83.9%, were achieved with $N = 300$ samples with Non-linear-param and Linear-param, respectively. These results indicate that Non-linear-param led to a relative decrease in error rate of 8.7% compared with the optimal Linear-param. According to the NIST matched-pairs sentence-segment word error test (MAPSSWE, Pallet et al. 1990) the difference is significant (p -value < 0.001 ; significance at p -value < 0.01). Because Non-linear-param provided better accuracy than Linear-param, and its optimal value of f_{th} is 80 Hz,

Table 2. Classification accuracy and sensitivity with non-linear-param ($N = 300$ samples, $\alpha = 30$, and $f_{th} = 80\text{Hz}$), linear-param and mel-cepstral.

Parametrization	Overall Classification Accuracy	Sensitivity (SEP 2)
Non-linear-param N = 300 samples, $\alpha = 30$, and $f_{th} = 80$ Hz (Fully-supervised-initialization)	85.3	86.3
Linear-param, N = 300 samples (Fully-supervised-initialization)	83.9	85.9
Linear-param, N = 400 samples (Fully-supervised-initialization)	82.8	84.8
Linear-param, N = 500 samples (Fully-supervised-initialization)	82.2	83.4
Mel-cepstral (Fully-supervised-initialization)	60.5	67.2
Spectrogram cross correlation	N.A.	60.6

a frequency compression above 80 Hz helped to better represent those acoustic events characterized by components above 80 Hz (e.g. HB and ship noise). It is worth mentioning that the frequency resolution of the Non-linear-param for frequencies less than f_{th} is equivalent to the frequency resolution of the Linear-param. When comparing Mel-cepstral with the other parametrizations, Mel-cepstral led to an increase in classification error of 169% (from 14.7% to 39.5%) and 145% (from 16.1% to 39.5%) relative to Non-linear-param and Linear-param, respectively. This strongly supports the proposed parametrization scheme and our strategy to re-think the frequency compression curve applied to the problem of whale vocalization detection and classification. It is worth mentioning that we had to run the experiments with Mel-cepstral on ‘Corcovado-Songmeter’ database because the sensitivities with respect to SEP2 obtained here are not comparable with the one achieved in Putland et al. (2018) that targets Bryde’s whale (*Balaenoptera edeni*) vocalizations based on only three categories of annotated event. In this study, 43 event types (Table 1) were annotated and modelled so we would expect to obtain higher detection and classification errors; also the target whale species here is not Bryde’s whale.

5.2. Noise modelling evaluation

Table 3 shows the average detection and classification accuracy with the two noise model topologies applied to the ‘Corcovado-Songmeter’ database described in Section 3, i.e. three-state left-to-right with state skip transition (Figure 5(b)) and three-state left-to-right without state skip transition (Figure 5(a)). The experiments were carried out with the tuned Non-linear-param as in Table 2. The HMM training procedure made use of the proposed Fully-supervised-initialization model initialization. According to Table 3, the noise model with state skip transition topology delivers a reduction in error rate equal to 3% relative when compared to the topology without state skip transition. According to

Table 3. Accuracy for different noise model topologies with the tuned non-linear-param using the proposed fully-supervised-initialization model initialization.

Noise topology	Accuracy
Three-states with state skip transition	85.3
Three-states without state skip transition	84.8

Table 4. Comparison of HMM training initialization. Non-linear-param as in Table 3 was employed in both cases.

GMM initialization	Accuracy
Fully-supervised-initialization	85.3
Semi-supervised-initialization	64.2

the NIST matched-pairs sentence-segment word error test (MAPSSWE, Pallet et al. 1990) the difference is significant (p -value < 0.001 ; significance taken at p -value < 0.01). This can probably be explained by the fact that the separation between two vocalizations is in some cases as short as one frame, and these silences are better represented when noise is modelled with a topology that includes the state skip transition.

5.3. Model initialization evaluation

Table 4 shows detection and classification accuracy of the two model initialization methods applied to the ‘Corcovado-Songmeter’ database (Section 4.2), i.e. Fully-supervised-initialization and Semi-supervised-initialization. Non-linear-param as in Table 3 was employed in both cases. Noise models used the left-to-right three-state with state skip transition topology (Figure 5(b)). According to Table 4, the proposed initialization method led to a dramatic reduction in error rate of 59% relative to the ordinary Semi-supervised-initialization provided by Kaldi. This may be a consequence of the stereotyped nature of the AA, SEP1 and SEP2 blue whale vocalizations. The Fully-supervised-initialization initialization described in Section 4.2 takes advantage of this fact, and therefore generates more representative and accurate initial models.

5.4. Validation experiments

Results of the validation experiments were obtained with the following configurations: System 1: Non-linear-param combined with Fully-supervised-initialization; System 2: Non-linear-param combined with Semi-supervised-initialization; and System 3: Mel-cepstral combined with Fully-supervised-initialization. The noise model was represented with the three-state left-to-right with state skip transition HMM topology (Figure 5(b)) for the three systems described. Detection of SEP2 phrases using Systems 1, 2, and 3 (where detection of song unit D or S2.3 events as proxy for the entire phrase) was compared with the average daily number of SEP2 published in (Buchan et al. 2015) and shown in Figure 10. Detection of SEP2 phrases using Systems 1, 2, and 3 was compared with the Buchan et al. (2015) reference (Figure 10). As seen in Figure 10, System 1 follows much more closely the reference than Systems 2 or 3. The slight difference between System 1 and the reference is hard to analyse because the latter results from a correction made with a constant estimated empirically (see Section 4). Regarding Mel-cepstral, Figure 10 supports the conclusion in Section 5.1 that these features do not discriminate among whale vocalizations as well as the Non-linear-param. This must be due to the fact that Mel filterbank was optimized for speech recognition and not for bioacoustics. To compare the performance of the three systems, the root mean square error (MSE) between each evaluated system and the reference was computed. The proposed detection

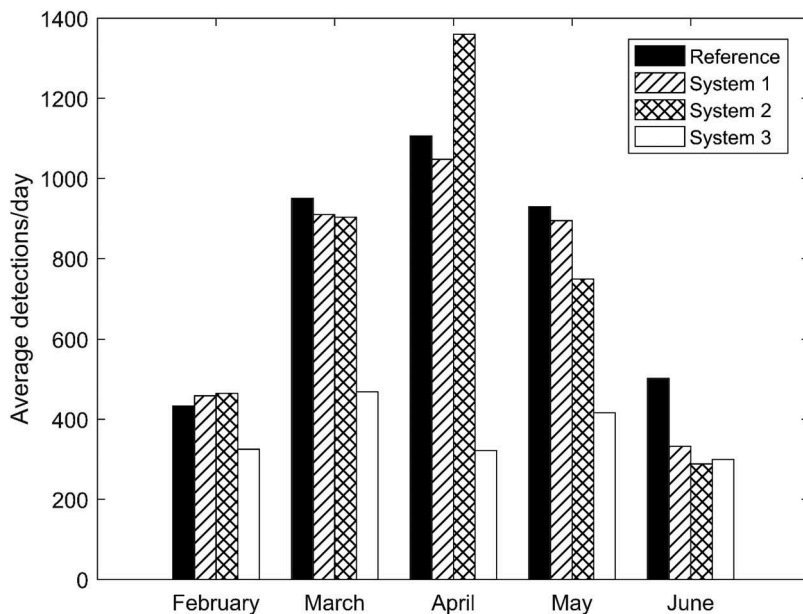


Figure 10. Results of average SEP2 song detections per day for the Corcovado-MARU database. ‘Reference’ bars represent the results reported in (Buchan et al. 2015). ‘System 1’ and ‘System 2’ bars represent the results with Non-linear-param combined with Fully-supervised-initialization and Semi-supervised-initialization, respectively. ‘System 3’ bars represent results with Mel-cepstral combined with Fullysupervised-initialization.

and classification system, System 1, led to an MSE of 82.0, which is much lower than the MSE with respect to the reference obtained with System 2 (MSE = 344.1) and System 3 (MSE = 440.2). In fact, System 2 and System 3 deliver MSE’s that are 320% and 437% greater than System 1. These results indicate that System 1 is the optimal system for whale vocalization classification and detection. This also confirms the pertinence of the frequency compression curve for processing whale vocalizations; the accurate generation of initial models for training estimated with a set of handmade annotated or labelled signals; and, the model scheme that includes simultaneous whale vocalizations.

5.5. Performance comparison with the spectrogram cross-correlation method

An important criterion when comparing the proposed HMM-based method to other methods, is the amount of human supervision time required and the computational processing time. Here, training procedure does involve human analyst time, however in principle, the training stage needs to be done only once with a database of annotated events targeted by the study in question. The time required to annotate the 157.5 hours of the ‘Corcovado-Songmeter’ database was approximately 43.4 analyst hours; and the training procedure using the annotated data demanded 1.3 hours of an Intel i7 desktop PC with 32 GB of RAM. Once the system was trained, it could be used to detect vocalizations in a different database without the need of human supervision. Then, we compared the time demanded to obtain the results of Figure 10 with our trained system and with the

spectrogram cross-correlation method. The time required to run our trained system on the five months of data ('Corcovado-MARU' database) was 5.0 hours with the same PC employed to train the system. In comparison, and disregarding the time required to run the spectrogram cross-correlation detector on five months of data, which depending on computational power can be hours or a day, the visual review of detections according to the method in Buchan et al. (2015) involved 17.3 analyst-days. Consequently, the proposed HMM-based automatic system allows the processing of huge amounts of data with no human-supervision aside from the model training phase and achieves similar accuracies to the commonly used spectrogram cross-correlation method, which cannot be done without human review of detection results.

5.6. Implications for monitoring baleen whales in the southeast pacific

The coast of Chile is host to 50% of the world's cetacean species (Aguayo-Lobo et al. 1998), most of which are currently classified as Vulnerable or Endangered (<http://www.iucnredlist.org/>) following commercial whaling. This study contributes to advancing the PAM of endangered whale populations off the coast of Chile and in the southeast Pacific region, which is an extensive area of ocean that remains poorly covered in terms of ocean observation (acoustic and non-acoustic) in general, and marine mammal monitoring in particular. Moreover, marine bioacoustics is a relatively new field in this region and research effort remains limited. For example, there are only six papers on blue whale acoustics in this region (Cumplings and Thompson 1971; Stafford et al. 1999a; Buchan et al. 2010, 2014, 2015; Buchan and Quiñones 2016), compared with over a dozen papers on blue whale acoustics in the North Pacific (Thompson 1965; Thompson et al. 1996; Rivers 1997; Stafford et al. 1998, 1999b, 2001, 2005, 2007, 2009; Thode et al. 2000; McDonald et al. 2001; Wiggins et al. 2005; Rankin et al. 2006; Oleson et al. 2007a, 2007b; Širović 2016). This is both a reflection of a low number of researchers working in this field and limited financial resources to collect and analyse PAM data. Given this scenario, developing methods that require a limited degree of human supervision, or even no human supervision, is particularly important to advance the analysis of existing datasets in this region, but also to demonstrate the feasibility of PAM studies to national authorities and funding bodies in Latin America who are generally unfamiliar with the broad applications of PAM technologies (compiled in Au and Lammers 2016). More efficient analytical methods also advance the possibility of real-time or near real-time PAM, that hold real promise for decision making, e.g. the real-time detection of endangered whale presence for planning human activities in coastal and offshore environments, or reducing the risk of fatal collisions between whales and large ships (Baumgartner et al. 2018).

Beyond the southeast Pacific, we hope that this system can be applied to other regions of the world ocean; similarly, we hope that this system can be applied to vocal cetacean species other than blue whales.

6. Conclusions

The contributions of this study are: the accurate modelling of whale vocalizations including overlapping vocalizations; proposing a frequency compression curve for

processing whale vocalizations; the generation of accurate initial models for training with an annotated database; using a state-of-the-art platform to run machine learning experiments; and, advancing methods for the acoustic monitoring of Endangered baleen whales off the coast of Chile and the southeast Pacific.

This study provides an automated system, without human intervention, for the detection and classification of single and overlapping blue whale vocalizations recorded off the coast of Chile (SEP1, SEP2, AA and D-calls) using HMM implemented with the Kaldi speech recognition toolkit, with 85.3% accuracy. This is the first automatic method for the detection and classification of blue whale vocalizations off Chile and can be applied in the future to other baleen whale vocalization types. To the best of our knowledge this is also the first time that Kaldi has been used for analysing whale vocalizations. In addition, this study proposes a new frequency compression curve for analysing whale vocalizations that improves detection and classification, which we recommend be used by other researchers that are processing low-frequency (<200 Hz) whale vocalizations. Consequently, another consequence of this study is the fact that the use of Mel cepstral features in bioacoustics in general should be revised and replaced by more ad-hoc parameters optimized with target species in mind. Finally, the proposed system in this study has been validated by reproducing very similar results without human supervision to published results obtained via spectrogram cross-correlation. Further training and testing to expand the repertoire of target signals of this system and further validation with other published reference datasets should be the focus of future research.

Notes

1. <http://www.iucnredlist.org/>.
2. <http://kaldi-asr.org/>.
3. <http://www.wildlifeacoustics.com/>.
4. <http://www.birds.cornell.edu/brp/>.

Acknowledgments

This research was primarily funded by the Office of Naval Research Global Grant number N62909-17-1-2013 (Universidad de Chile). This work would not have been possible without funding from the COPAS Sur-Austral Center (Universidad de Concepción) through CONICYT grants PFB31 and AFB170006 for the collection of the Songmeter passive acoustic data and support to Dr Susannah Buchan. We thank Dr. Ivan Perez-Santos (Centro iMar, Universidad de Los Lagos) for his help during the collection of this data. Our thanks also go to Dr. Rodrigo Hucke-Gaete (Universidad Austral de Chile and Centro Ballena Azul) for the collection of the MARU dataset. Dr. Susannah Buchan was also supported during the writing of this manuscript by CONICYT grant R16A10003 to Centro de Estudios Avanzados en Zonas Aridas and Office of Naval Research Global Grant number N00014-17-2606.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Comisión Nacional de Investigación Científica y Tecnológica [AFB170006, PFB31, R16A10003]; Office of Naval Research Global [N00014-17-2606, N62909-17-1-2013].

References

- Agranat I. 2013. Bat species identification from zero crossing and full spectrum echolocation calls using hidden Markov models, fisher scores, unsupervised clustering and balanced winnow pairwise classifiers. ICA2013. Proceedings of Meetings on Acoustics; Jun 2–7; Montreal, Canada. p. 010016.
- Aguayo-Lobo A, Torres D, Acevedo J. 1998. Los Mamíferos Marinos de Chile: 1. Cetacea [The marine mammals of Chile: 1. Cetacea]. Ser Cient INACH. 48:19–159. Spanish.
- Au WW, Lammers MO. 2016. Listening in the ocean. New York (NY): Springer.
- Au WW, Pack AA, Lammers MO, Herman LM, Deakos MH, Andrews K. 2006. Acoustic properties of humpback whale songs. *J Acoust Soc Am.* 120(2):1103–1110.
- Balcazar NE, Klinck H, Nieuwkirk SL, Mellinger DK, Klinck K, Dziak RP, Rogers TL. 2017. Using calls as an indicator for Antarctic blue whale occurrence and distribution across the southwest Pacific and southeast Indian Oceans. *Mar Mammal Sci.* 33(1):172–186.
- Baumgartner MF, Stafford KM, Latha G. 2018. Near real-time underwater passive acoustic monitoring of natural and anthropogenic sounds. In: Venkatesan R, Tandon A, D’Asaro E, Atmanand M, editors. *Observing the oceans in real time.* Cham: Springer; p. 203–226.
- Bejder M, Johnston DW, Smith J, Friedlaender A, Bejder L. 2016. Embracing conservation success of recovering humpback whale populations: evaluating the case for downlisting their conservation status in Australia?. *Mar Policy.* 66:137–141.
- Bioacoustics Research Program. 2012. Raven Pro: interactive sound analysis software (version 1.5) [computer software]. [Internet]. Ithaca (NY): The Cornell Lab of Ornithology; [accessed 2018 Apr 2]. <http://www.birds.cornell.edu/raven>.
- Branch TA, Stafford K, Palacios D, Allison C, Bannister J, Burton C, Cabrera E, Carlson C, Galletti Vernazzani B, Gill PC, et al. 2007. Past and present distribution, densities and movements of blue whales *Balaenoptera musculus* in the Southern Hemisphere and northern Indian Ocean. *Mammal Rev.* 37(2):116–175.
- Brown JC, Smaragdis P. 2009. Hidden Markov and Gaussian mixture models for automatic call classification. *J Acoust Soc Am.* 125(6):EL221–EL224.
- Buchan SJ, Hucke-Gaete R, Rendell L, Stafford KM. 2014. A new song recorded from blue whales in the Corcovado Gulf, Southern Chile, and an acoustic link to the Eastern Tropical Pacific. *Endanger Species Res.* 23(3):241–252.
- Buchan SJ, Hucke-Gaete R, Stafford KM, Clark CW. 2018. Occasional acoustic presence of Antarctic blue whales on a feeding ground in southern Chile. *Mar Mammal Sci.* 34(1):220–228.
- Buchan SJ, Quiñones RA. 2016. First insights into the oceanographic characteristics of a blue whale feeding ground in northern Patagonia, Chile. *Mar Ecol Prog Ser.* 554:183–199.
- Buchan SJ, Rendell LE, Hucke-Gaete R. 2010. Preliminary recordings of blue whale (*Balaenoptera musculus*) vocalizations in the Gulf of Corcovado, northern Patagonia, Chile. *Mar Mammal Sci.* 26(2):451–459.
- Buchan SJ, Stafford KM, Hucke-Gaete R. 2015. Seasonal occurrence of southeast Pacific blue whale songs in southern Chile and the eastern tropical Pacific. *Mar Mammal Sci.* 31(2):440–458.
- Clark CW, Ellison WT, Southall BL, Hatch L, Van Parijs SM, Frankel A, Ponirakis D. 2009. Acoustic masking in marine ecosystems: intuitions, analysis, and implication. *Mar Ecol Prog Ser.* 395:201–222.
- Croll DA, Marinovic B, Benson S, Chavez FP, Black N, Ternullo R, Tershy BR. 2005. From wind to whales: trophic links in a coastal upwelling system. *Mar Ecol Prog Ser.* 289:117–130.

- Cummings WC, Thompson PO. 1971. Underwater sounds from the blue whale, *Balaenoptera musculus*. *J Acoust Soc Am.* 50(4B):1193–1198.
- Davis GE, Baumgartner MF, Bonnell JM, Bell J, Berchok C, Thornton JB, Brault S, Buchanan G, Charif RA, Cholewiak D, et al. 2017. Long-term passive acoustic recordings track the changing distribution of North Atlantic right whales (*Eubalaena glacialis*) from 2004 to 2014. *Sci Rep.* 7(1):13460.
- Delarue J, Todd SK, Van Parijs SM, Di Iorio L. 2009. Geographic variation in Northwest Atlantic fin whale (*Balaenoptera physalus*) song: implications for stock structure assessment. *J Acoust Soc Am.* 125(3):1774–1782.
- Dugan PJ, Rice AN, Urazghildiiev IR, Clark CW. 2010. North Atlantic right whale acoustic signal processing: part I. Comparison of machine learning recognition algorithms. Proceedings of the 2010 Long Island Systems, Applications and Technology Conference; May 7; Long Island. p. 1–6.
- Dziak RP, Bohnenstiehl DR, Stafford KM, Matsumoto H, Park M, Lee WS, Fowler MJ, Lau TK, Haxel JH, Mellinger DK. 2015. Sources and levels of ambient ocean sound near the Antarctic Peninsula. *PLoS One.* 10(4):e0123425.
- Erbe C, Verma A, McCauley R, Gavrilov A, Parnum I. 2015. The marine soundscape of the Perth Canyon. *Prog Oceanogr.* 137:38–51.
- Español-Jiménez S, van der Schaar M. 2018. First record of humpback whale songs in Southern Chile: analysis of seasonal and diel variation. *Mar Mammal Sci.* 34(3):718–733.
- Espósito AM, D'Auria L, Giudicepietro F, Peluso R, Martini M. 2013. Automatic recognition of landslides based on neural network analysis of seismic signals: an application to the monitoring of Stromboli volcano (southern Italy). *Pure Appl Geophys.* 170(11):1821–1832.
- Galletti-Vernazzani B, Jackson JA, Cabrera E, Carlson CA, Brownell RL Jr. 2017. Estimates of abundance and trend of Chilean blue whales off Isla de Chiloé, Chile. *PLoS One.* 12(1):e0168646.
- Gavrilov AN, McCauley RD, Gedamke J. 2012. Steady inter and intra-annual decrease in the vocalization frequency of Antarctic blue whales. *J Acoust Soc Am.* 131(6):4476–4480.
- Gold B, Morgan N, Ellis D. 2011. Speech and audio signal processing: processing and perception of speech and music. Hoboken (NJ): John Wiley & Sons.
- Hatch LT, Clark CW, Van Parijs SM, Frankel AS, Ponirakis DW. 2012. Quantifying loss of acoustic communication space for right whales in and around a US National Marine Sanctuary. *Conserv Biol.* 26(6):983–994.
- Helble TA, Ierley GR, D'Spain GL, Martin SW. 2015. Automated acoustic localization and call association for vocalizing humpback whales on the Navy's Pacific Missile Range Facility. *J Acoust Soc Am.* 137(1):11–21.
- Hildebrand JA. 2009. Anthropogenic and natural sources of ambient noise in the ocean. *Mar Ecol Prog Ser.* 395:5–20.
- Huang XD, Ariki Y, Jack MA. 1990. Hidden Markov models for speech recognition. New York (NY): Columbia University Press.
- Hucke-Gaete R, Álvarez R, Navarro M, Ruiz J, Lo Moro P. 2010. Investigación para desarrollo de área marina costera protegida Chiloé-Palena-Guaitecas [Research for development of protected coastal marine area Chiloé-Palena-Guaitecas]. Chile: Gobierno Regional de Los Lagos. Final report FNDR-BID TURISMO Cód. BIP No. 30040215–0. Spanish.
- Hucke-Gaete R, Osman LP, Moreno CA, Findlay KP, Ljungblad DK. 2004. Discovery of a blue whale feeding and nursing ground in southern Chile. *Proc R Soc Lond B Biol Sci.* 271(Suppl 4):S170–S173.
- Knowlton AR, Hamilton PK, Marx MK, Pettis HM, Kraus SD. 2012. Monitoring North Atlantic right whale *Eubalaena glacialis* entanglement rates: a 30 yr retrospective. *Mar Ecol Prog Ser.* 466:293–302.
- Laist DW, Knowlton AR, Mead JG, Collet AS, Podesta M. 2001. Collisions between ships and whales. *Mar Mammal Sci.* 17(1):35–75.
- Li J, Deng L, Haeb-Umbach R, Gong Y. 2015. Robust automatic speech recognition: a bridge to practical applications. Waltham (MA): Academic Press.
- Ljungblad DK, Clark CW, Shimada H. 1998. A comparison of sounds attributed to pygmy blue whales (*Balaenoptera musculus breviceauda*) recorded south of the Madagascar Plateau and

- those attributed to 'true' blue whales (*Balaenoptera musculus*) recorded off Antarctica. Report-International Whaling Commission. 48:439–442.
- McDonald MA, Calambokidis J, Teranishi AM, Hildebrand JA. 2001. The acoustic calls of blue whales off California with gender data. *J Acoust Soc Am.* 109(4):1728–1735.
- McDonald MA, Hildebrand JA, Mesnick S. 2009. Worldwide decline in tonal frequencies of blue whale songs. *Endanger Species Res.* 9(1):13–21.
- McDonald MA, Mesnick SL, Hildebrand JA. 2006. Biogeographic characterization of blue whale song worldwide: using song to identify populations. *J Cetac Res Manage.* 8:55–65.
- McKenna M, Katz S, Wiggins S, Ross D, Hildebrand J. 2012. A quieting ocean: unintended consequence of a fluctuating economy. *J Acoust Soc Am.* 132(3):EL169–EL175.
- Mellinger DK, Clark CW. 2000. Recognizing transient low-frequency whale sounds by spectrogram correlation. *J Acoust Soc Am.* 107(6):3518–3529.
- Mellinger DK, Stafford KM, Moore SE, Dziak RP, Matsumoto H. 2007. An overview of fixed passive acoustic observation methods for cetaceans. *Oceanography.* 20(4):36–45.
- Moore SE. 2008. Marine mammals as ecosystem sentinels. *J Mammal.* 89(3):534–540.
- Neilson JL, Gabriele CM, Jensen AS, Jackson K, Straley JM. 2012. Summary of reported whale–vessel collisions in Alaskan waters. *J Mar Biol.* 2012:1–18.
- Nieukirk SL, Fregosi S, Mellinger DK, Klinck H. 2016. A complex baleen whale call recorded in the Mariana Trench Marine National Monument. *J Acoust Soc Am.* 140(3):EL274–EL279.
- Noda JJ, Travieso CM, Sánchez-Rodríguez D. 2016. Methodology for automatic bioacoustic classification of anurans based on feature fusion. *Expert Syst Appl.* 50:100–106.
- Oleson EM, Calambokidis J, Burgess WC, McDonald MA, LeDuc CA, Hildebrand JA. 2007a. Behavioral context of call production by eastern North Pacific blue whales. *Mar Ecol Prog Ser.* 330:269–284.
- Oleson EM, Wiggins S, Hildebrand JA. 2007b. Temporal separation of blue whale call types on a southern California feeding ground. *Anim Behav.* 74:881–894.
- Pallet DS, Fisher WM, Fiscus JG. 1990. Tools for the analysis of benchmark speech recognition tests. Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing; Apr 3–6; Albuquerque, NM. p. 97–100.
- Potamitis I, Ntalampiras S, Jahn O, Riede K. 2014. Automatic bird sound detection in long real-field recordings: applications and tools. *Appl Acoust.* 80:1–9.
- Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, et al. 2011. The Kaldi speech recognition toolkit. In IEEE 2011 workshop on automatic speech recognition and understanding; Dec 11–15. Big Island (HI): IEEE Signal Processing Society.
- Putland R, Ranjard L, Constantine R, Radford C. 2018. A hidden Markov model approach to indicate Bryde's whale acoustics. *Ecol Indic.* 84:479–487.
- Rabiner LR. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE.* 77(2):257–286.
- Rabiner LR, Juang BH. 1986. An introduction to hidden Markov models. *IEEE ASSP Mag.* 3(1):4–16.
- Rabiner LR, Juang BH. 1993. Fundamentals of speech recognition. Englewood Cliffs (NJ): PTR Prentice Hall.
- Ranjard L, Reed BS, Landers TJ, Rayner MJ, Friesen MR, Sagar RL, Dunphy BJ. 2017. MatlabHTK: a simple interface for bioacoustic analyses using hidden Markov models. *Methods Ecol Evol.* 8(5):615–621.
- Ranjard L, Withers SJ, Brunton DH, Ross HA, Parsons S. 2015. Integration over song classification replicates: song variant analysis in the hihi. *J Acoust Soc Am.* 137(5):2542–2551.
- Rankin S, Barlow J, Stafford KM. 2006. Blue whale (*Balaenoptera musculus*) sightings and recordings south of the Aleutian Islands. *Mar Mammal Sci.* 22(3):708–713.
- Read AJ, Drinker P, Northridge S. 2006. Bycatch of marine mammals in US and global fisheries. *Conserv Biol.* 20(1):163–169.

- Redfern J, McKenna M, Moore T, Calambokidis J, Deangelis M, Becker E, Barlow J, Forney K, Fiedler P, Chivers S. 2013. Assessing the risk of ships striking large whales in marine spatial planning. *Conserv Biol.* 27(2):292–302.
- Redfern JV, Hatch LT, Caldow C, DeAngelis ML, Gedamke J, Hastings S, Henderson L, McKenna MF, Moore TJ, Porter MB. 2017. Assessing the risk of chronic shipping noise to baleen whales off Southern California, USA. *Endanger Species Res.* 32:53–167.
- Ren Y, Johnson MT, Clemins PJ, Darre M, Glaeser SS, Osiejuk TS, Out-Nyarko E. 2009. A framework for bioacoustic vocalization analysis using hidden Markov models. *Algorithms.* 2(4):1410–1428.
- Rivers JA. 1997. Blue whale, *Balaenoptera musculus*, vocalizations from the waters off central California. *Mar Mammal Sci.* 13(2):186–195.
- Rocha RC, Clapham PJ, Ivashchenko YV. 2014. Emptying the oceans: a summary of industrial whaling catches in the 20th century. *Mar Fish Rev.* 76(4):37–48.
- Rolland RM, Parks SE, Hunt KE, Castellote M, Corkeron PJ, Nowacek DP, Wasser SK, Kraus SD. 2012. Evidence that ship noise increases stress in right whales. *Proc R Soc Lond B Biol Sci.* 279(1737):2363–2368.
- Samaran F, Stafford KM, Branch TA, Gedamke J, Royer JY, Dziak RP, Guinet C. 2013. Seasonal and geographic variation of southern blue whale subspecies in the Indian Ocean. *PLoS One.* 8(8):e71561.
- Scheifele PM, Johnson MT, Fry M, Hamel B, Laclede K. 2015. Vocal classification of vocalizations of a pair of Asian small-clawed otters to determine stress. *J Acoust Soc Am.* 138(1):EL105–EL109.
- Shamir L, Yerby C, Simpson R, von Benda-Beckmann AM, Tyack P, Samarra F, Miller P, Wallin J. 2014. Classification of large acoustic datasets using machine learning and crowdsourcing: application to whale calls. *J Acoust Soc Am.* 135(2):953–962.
- Širović A. 2016. Variability in the performance of the spectrogram correlation detector for North-east Pacific blue whale calls. *Bioacoustics.* 25(2):145–160.
- Širović A, Hildebrand JA, Wiggins SM. 2007. Blue and fin whale call source levels and propagation range in the Southern Ocean. *J Acoust Soc Am.* 122(2):1208–1215.
- Širović A, Hildebrand JA, Wiggins SM, McDonald MA, Moore SE, Thiele D. 2004. Seasonality of blue and fin whale calls and the influence of sea ice in the Western Antarctic Peninsula. *Deep Sea Res Part 2 Top Stud Oceanogr.* 51(17–19):2327–2344.
- Sousa-Lima RS, Engel MH, Sábato V, Lima BR, Queiróz TS, Brito MR, Fernandes DP, Martins CA, Hatum PS, Casagrande T, et al. 2018. Acoustic ecology of humpback whales in Brazilian waters investigated with basic and sophisticated passive acoustic technologies over 17 years. *West Indian Ocean J Mar Sci. Special Issue 1:* 23–40.
- Spillmann B, van Schaik CP, Setia TM, Sadjadi SO. 2017. Who shall I say is calling? Validation of a caller recognition procedure in Bornean flanged male orangutan (*Pongo pygmaeus wurmbii*) long calls. *Bioacoustics.* 26(2):109–120.
- Stafford KM, Chapp E, Bohnenstiel DR, Tolstoy M. 2011. Seasonal detection of three types of “pygmy” blue whale calls in the Indian Ocean. *Mar Mammal Sci.* 27(4):828–840.
- Stafford KM, Citta JJ, Moore SE, Daher MA, George JE. 2009. Environmental correlates of blue and fin whale call detections in the North Pacific Ocean from 1997 to 2002. *Mar Ecol Prog Ser.* 395:37–53.
- Stafford KM, Fox CG, Clark DS. 1998. Long-range acoustic detection and localization of blue whale calls in the northeast Pacific Ocean. *J Acoust Soc Am.* 104(6):3616–3625.
- Stafford KM, Mellinger DK, Moore SE, Fox CG. 2007. Seasonal variability and detection range modeling of baleen whale calls in the Gulf of Alaska, 1999–2002. *J Acoust Soc Am.* 122(6):3378–3390.
- Stafford KM, Moore SE, Fox CG. 2005. Diel variation in blue whale calls recorded in the eastern tropical Pacific. *Anim Behav.* 69(4):951–958.
- Stafford KM, Nieukirk SL, Fox CG. 1999a. Low-frequency whale sounds recorded on hydrophones moored in the eastern tropical Pacific. *J Acoust Soc Am.* 106(6):3687–3698.

- Stafford KM, Nieukirk SL, Fox CG. 1999b. An acoustic link between blue whales in the eastern tropical pacific and the Northeast Pacific. *Mar Mamm Sci.* 15(4):1258–1268.
- Stafford KM, Nieukirk SL, Fox CG. 2001. Geographic and seasonal variation of blue whale calls in the North Pacific. *J Cetac Res Manage.* 3(1):65–76.
- Stevens SS, Volkman J. 1940. The relation of pitch to frequency: a revised scale. *Am J Psychol.* 53(3):329–353.
- Stimpert AK, Au WW, Parks SE, Hurst T, Wiley DN. 2011. Common humpback whale (*Megaptera novaeangliae*) sound types for passive acoustic monitoring. *J Acoust Soc Am.* 129(1):476–482.
- Thode AM, D’Spain GL, Kuperman WA. 2000. Matched-field processing, geoacoustic inversion, and source signature recovery of blue whale vocalizations. *J Acoust Soc Am.* 107(3):1286–1300.
- Thomisch K, Boebel O, Clark CW, Hagen W, Spiesecke S, Zitterbart DP, Van Opzeeland I. 2016. Spatio-temporal patterns in acoustic presence and distribution of Antarctic blue whales *Balaenoptera musculus intermedia* in the Weddell Sea. *Endanger Species Res.* 30:239–253.
- Thompson PO. 1965. Marine biological sounds west of San Clemente Island: diurnal distributions and effects of ambient noise during July 1963. San Diego (CA): US Navy Electronics Laboratory. Report No.: NEL-1290.
- Thompson PO, Cummings WC, Ha SJ. 1986. Sounds, source levels, and associated behavior of humpback whales, Southeast Alaska. *J Acoust Soc Am.* 80(3):735–740.
- Thompson PO, Findley LT, Vidal O, Cummings WC. 1996. Underwater sounds of blue whales, *Balaenoptera musculus*, in the Gulf of California, Mexico. *Mar Mamm Sci.* 12(2):288–293.
- Tripovich JS, Klinck H, Nieukirk SL, Adams T, Mellinger DK, Balcazar NE, Klinck K, Hall EJ, Rogers TL. 2015. Temporal segregation of the Australian and Antarctic blue whale call types (*Balaenoptera musculus* spp.). *J Mammal.* 96(3):603–610.
- Trites AW, Christensen V, Pauly D. 1997. Competition between fisheries and marine mammals for prey and primary production in the Pacific Ocean. *J Northwest Atl Fish Sci.* 22:173–187.
- Van Opzeeland I, Boebel O. 2018. Marine soundscape planning: seeking acoustic niches for anthropogenic sound. *J Ecoacoust.* 2:5. GSNT.
- Van Parijs SM, Clark CW, Sousa-Lima RS, Parks SE, Rankin S, Risch D, Van Opzeeland IC. 2009. Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales. *Mar Ecol Prog Ser.* 395:21–36.
- Vanderlaan AS, Taggart CT. 2007. Vessel collisions with whales: the probability of lethal injury based on vessel speed. *Mar Mammal Sci.* 23(1):144–156.
- Viddi FA, Hucke-Gaete R, Torres-Florez JP, Ribeiro S. 2010. Spatial and seasonal variability in cetacean distribution in the fjords of northern Patagonia, Chile. *ICES J Mar Sci.* 67(5):959–970.
- Viterbi A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inf Theory.* 13(2):260–269.
- Wiggins SM, Oleson EM, McDonald MA, Hildebrand JA. 2005. Blue whale (*Balaenoptera musculus*) diel calling patterns offshore of Southern California. *Aquat Mammal.* 31(2):161–168.
- Williams R, Erbe C, Ashe E, Clark CW. 2015. Quiet(er) marine protected areas. *Mar Pollut Bull.* 100(1):154–161.