A DEEP LEARNING FRAMEWORK FOR CLASSIFYING SOUNDS OF MYSTICETE WHALES

22

Stavros Ntalampiras
Politecnico di Milano, Milan, Italy

22.1 INTRODUCTION

Bioacoustic signal processing has attracted a lot of attention during the last decade as it is able to offer robust solutions to problems with diverse needs [7,17,19]. The ultimate goal of frameworks processing bioacoustic signals is to provide a complete and accurate picture of the biodiversity of the habitat of interest toward its conservation [15]. Without such automatic frameworks the monitoring process is accomplished by human experts by thorough observation of the recorded data. Even though the quality of the work done by a human expert is superior to the services offered by a machine, there are many drawbacks with respect to monitoring carried out by humans: (a) they require more time as an algorithm is able to run faster than real-time, (b) the needed expeditions are costly or even impossible due to dangerous, inaccessible areas, (c) they are able to analyze a limited number of habitats, (d) they may also interfere with the behavior of the species of interest and alter its behavior.

An application domain falling under the umbrella of bioacoustic signal processing deals with the automatic categorization of marine mammal sounds. It comes out from the related literature that the specific domain is still not well explored with respect to others, such as processing of bird callings [18,20]. Mainly this is due to the fact that underwater sound recording requires more sophisticated equipment and resources in general. However the recent technological advancements in automatic recording units have facilitated the capturing of underwater sounds, thus nowadays one may easily have access to vast amount of the associated audio signals. These databases can be used for the development of automated methods which achieve biodiversity monitoring toward a better analysis of underwater life.

This study addresses the problem of classifying sounds coming from mysticetes based on the hypothesis the mammalian cortex uses a form of hierarchical decomposition for processing sound stimuli [24,25]. This results in a consistent distribution of the energy produced by each audio signal with respect to specific parts of the spectrum. Our aim is to design acoustic features able to capture this distribution and subsequently model them (and/or their evolution in time) for its automatic classification. The mysticete species included in the present study are: (a) *Blue whales*, (b) *Bowhead whales*, (c) *Fin whales*, (d) *Humpback whales*, and (e) *Southern Right whales*.

The problem is quite challenging since the related signals may exhibit similar temporal and spectral characteristics. Thus one must search for features able to capture even slight differences among the signals belonging to the above-mentioned species. Another issue is that the data of a specific class may exhibit distribution with varying characteristics mainly due to the noise coexisting with the signals of interest.

Here we use information coming both from cepstral and wavelet domains. Subsequently they are modeled by a method exploiting a discriminative classifier based on deep learning. The pattern modeling technique is adaptive while taking into account the following issues: (a) limited data associated with one or more classes, and (b) data set exhibiting imbalances with respect to one or more classes, i.e. data quantities among the classes are unequal. We followed a thorough experimental procedure using a publicly available data set and reached quite encouraging classification rates.

The rest of this article is organized as follows. Section 22.2 provides an overview of the related literature. Section 22.4 analyzes the modules which comprise the proposed classification framework with special attention to the universal background and reservoir modeling. The next section examines the capabilities of the proposed approach in a thorough and concise way. Finally Section 22.6 offers our conclusions as well as ideas for future work.

22.2 RELATED LITERATURE

Processing of sounds coming from mysticete and/or odontocete species has attracted the interest of the audio signal processing community quite recently. Halkias et al. [5] designed a method able to classify mysticete sounds of five species (Blue whale, Bowhead whale, Fin whale, Humpback whale, and Southern Right whale) under the presence of noise (ambient noise, mechanical noise, other species). Their method is based on Restricted Boltzmann Machine and Sparse Auto-Encoder fed on spectrogram ROIs while providing a recognition accuracy of 69% and 80% with and without the presence of noise respectively. Three time-frequency methods for recognizing fin and blue whale calls are presented in [12]. The methods include spectrogram matching, dynamic time wrapping, and vector quantization while the latter two operate on the frequency contour. The data set was recorded by the authors while they emphasize the strong and weak points of each method.

Study [23] processes sounds produced by ten killer whales and eight pilot whales close to the coasts of Norway, Iceland, and the Bahamas (Whale FM project). They were automatically analyzed and the killer whales were classified as Icelandic or Norwegian while the pilot ones were separated into Norwegian long-finned and Bahamas' short-finned pilot whales. The audio features are extracted out of the spectrogram while the classification is based on a distance metric weighted by Fisher discriminant scores. It is interesting to note that the proposed method performed better than the analysis of the citizens. In [1] the authors used the short-time Fourier and wavelet packet transforms along with a multi-layer perceptron (MLP) to analyze blue whale calls. The proposed system is able to classify the vocalizations into A, B, and D blue whale classes.

Paper [13] employs two types of neural networks (based on competitive learning and Kohonen feature mapping respectively) in order to analyze the repertoire of false killer whale vocalizations. The authors used duty cycle measurements and peak frequency as signal characteristics while three major categories were discovered: ascending whistles, low-frequency pulse trains, and high-frequency pulse

trains. It should be noted that the vocalizations were captured by two false killer whales, one male and one female, located at Sea Life Park, Oahu, Hawaii.

Brown and Miller [3] applied four Dynamic Time Wrapping algorithms on a set of calls by Northern Resident whales which may be categorized into seven different classes. Their features included the low-frequency contour, the high-frequency contour, their derivatives, and weighted sums of the distances corresponding to LFC with HFC, LFC with its derivative, and HFC with its derivative. Subsequently Brown and Smaragdis [4] used hidden Markov models (HMMs) and Gaussian mixture models (GMMs) to classify seven types of calls coming from Northern Resident killer whales. Their feature set was a time-frequency decomposition of the recorded signals.

An interesting approach is presented in [11], based on spectrogram correlation. The corpus consisted of bowhead whale's (Balaena mysticetus) end notes from songs recorded in Alaska in 1986 and 1988 while the method outperformed three other methods (matched filters, neural networks, and hidden Markov models).

Roch et al. [10] explain a method for classification of free-ranging delphinid vocalizations. The feature extraction concerned cepstral vectors associated with multisecond segments. The authors trained one Gaussian mixture model for each of the following three species: short-beaked and long-beaked common (Delphinus delphis and Delphinus capensis), Pacific white-sided (Lagenorhynchus obliquidens), and bottlenose (Tursiops truncates). Last but not least Wilcock [28] followed a fundamentally different approach and performed tracking of fin whales in the northeastern Pacific Ocean using measurements coming from a seafloor seismic network.

To the best of our knowledge there are no approaches in the literature exploiting a deep learning classifier in combination with a multidomain set of features for the classification of five Mysticete species.

22.3 ACOUSTIC FEATURES

This section explains the parameterization of the audio signals coming from the five whale species. Both frequency and wavelet domains were employed toward obtaining a spherical picture of the involved sound events. For convenience, the features extracted out of a Blue whale sound event are depicted in Fig. 22.1.

22.3.1 FREQUENCY DOMAIN FEATURES

The first feature set exploits the spectrogram of the sound event since it may reveal important information for its characterization. The Fast Fourier Transform is used while the signal is windowized in order to minimize the effect of spectral leakages, i.e. diminish the finite length sequence at the ends aiming at a periodic structure without discontinuities. There exists a gamut of window functions with very different spectral properties, e.g. main lobe widths and side lobe amplitudes [6]. Here we have employed the following windowing techniques to reduce edge effects in the FFT: (a) Blackman, (b) Hamming, (c) Hanning, and (d) rectangular.

Since classification of whale sounds is a relatively new task for the audio signal processing community and a standard windowing technique has not been established, we performed a series of examinations to determine the optimal one. Early experimentations showed that Blackman window-
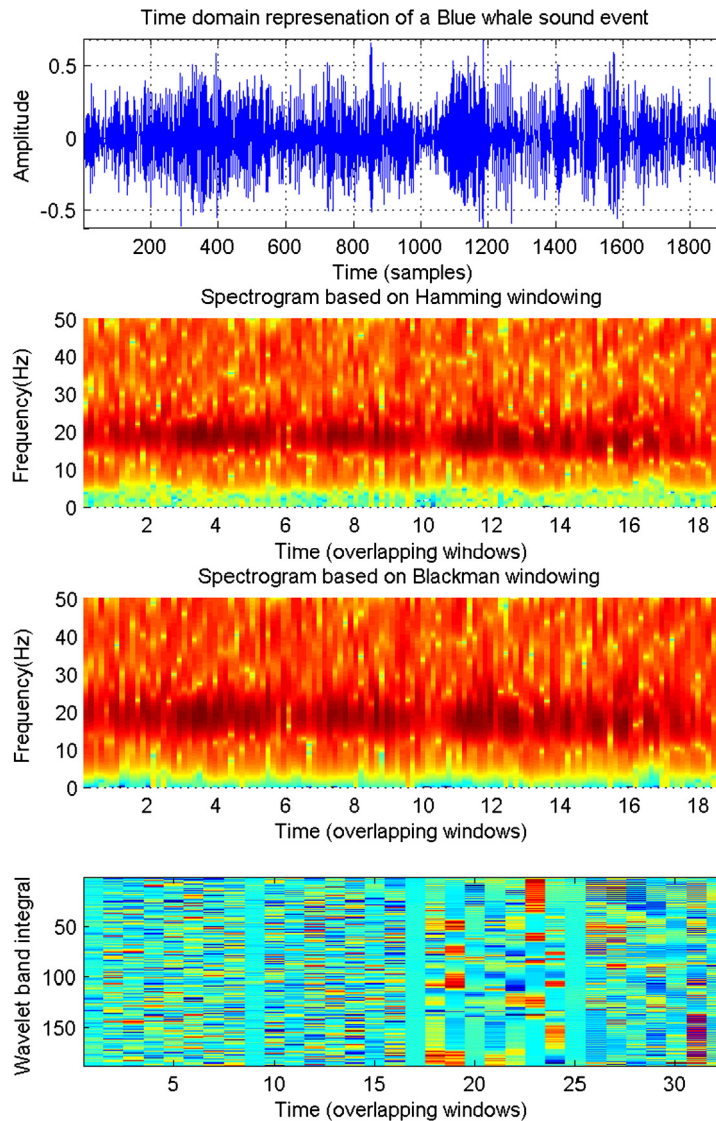
FIGURE 22.1

A representation of the feature sets used in this work. Both frequency and wavelet domains are considered.

ing offers the best spectral representation with respect to classification accuracy, thus it is favored in this work. As one may see in Fig. 22.1 the high energy parts of the spectrum are more emphasized while using Blackman window type than the Hamming one (Hamming was chosen since it is com-
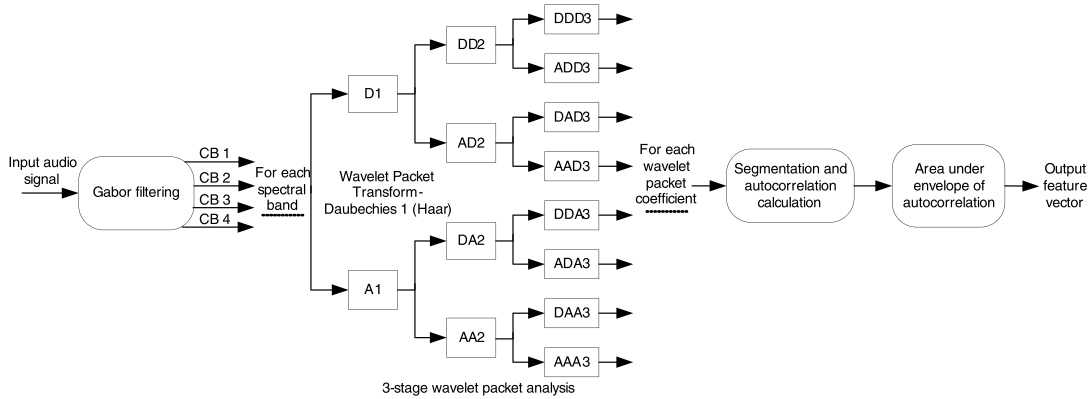
FIGURE 22.2

The block diagram of the process extracting the wavelet packet integration feature set.

monly used in audio processing applications). Thus the final feature vector is the energies of the short time Fourier transform after it is Blackman windowed.

22.3.2 WAVELET DOMAIN FEATURES

This group is extracted after a critical band-based multiresolution analysis of the signal takes place. Wavelets have become a common tool in many signal processing applications (bioacoustic signal enhancement [21], audio fingerprinting [2], speech/music discrimination [14] etc.). The uniqueness of the wavelet transform comes from its ability for processing time series, which include non-stationary power at many different frequencies. While the Fourier transform is based on smooth and predictable sinusoid functions, wavelets tend to be irregular and asymmetric. It is a dynamic windowing technique processing low and high frequency information content with different levels of analysis.

Wavelet packet (WP) analysis breaks up the signal and transforms it into shifted and scaled variants of the original (or mother) function. In this article we employed the Daubechies 1 (or Haar) function. The proposed methodology applies the discrete wavelet transform three subsequent times which is equivalent to a three-stage filtering while we retain both low and high frequency content. The feature extraction process is depicted in Fig. 22.2.

With this set we wish to obtain a vector with a complete analysis of the audio signal across different spectral areas while they are approximated by WP. This set takes into account that not all parts of the spectrum contain valuable information while some parts are highly contaminated with noise. After manual inspection of the recordings, we employed a filterbank with the frequency ranges denoted in Table 22.1 using Gabor bandpass filters based on a Gaussian kernel. Subsequently we extract three-level wavelet packets out of each spectral band while applying downsampling as Nyquist theorem requests, in order not to end up having the double amount of data. During the next stage we compute the autocorrelation envelope area with respect to each segmented wavelet coefficient and we normalize it by half the segment size. Finally we form a vector comprised of N normalized integration parameters, where N is the total number of the frequency bands multiplied by the number of the wavelet coefficients

Table 22.1 The Frequency Limits of the Wavelet Packet Integration Analysis			
Band Number	**Lower (Hz)**	**Center (Hz)**	**Upper (Hz)**
1	1	5	10
2	10	15	20
3	20	25	30
4	30	35	40
5	40	45	50

$(5 \times 8 = 40)$. This is the WP-integration feature vector and the block diagram for its computation is demonstrated in Fig. 22.2. They capture the variations exhibited by each wavelet coefficient within a group of predefined frequency bands. The normalized autocorrelation envelope area was chosen as the whale signals show differences in the content of the frequency bands we utilized.

22.4 THE CLASSIFICATION FRAMEWORK

We decided to apply a classification methodology approaching the problem from the following perspective: the Reservoir Network (RN) tries to determine the hyperplanes which separate the feature space while projecting them to a multidimensional space. They basically comprise recurrent neural networks, i.e., a deep learning architecture whose their main purpose is to capture the characteristics of high-level abstractions existing in the acquired data while designing multiple processing layers of complicated formations, i.e. non-linear functions. The advantage of RN is that the calculations involved in its readout layer are linear, thus of limited computational complexity and relatively small duration of the training process. Reservoir computing argues that since back propagation is computationally complex but typically does not influence the internal layers severely, it may be totally excluded from the training process. On the contrary, the readout layer is a generalized linear classification/regression problem associated with low complexity. In addition any potential network instability is avoided by enforcing a simple constraint on the random parameters of the internal layers.

In the following we provide a brief description of the RN.

22.4.1 RESERVOIR NETWORK

The trend in acoustic modeling suggests the usage of Reservoir Computing (RC) techniques [26]. An RN comprises an *a priori* fixed Recurrent Neural Network (RNN), the output layer of which is linear. An RN, whose topology is depicted in Fig. 22.3, includes neurons with non-linear activation functions which are connected to the inputs (input connections) and to each other (recurrent connections). These two types of connections have randomly generated weights, which are kept fixed during both the training and operational phase. Finally, a linear function is associated with each output node.

Its parameters are the weights of the output connections and are trained to achieve a specific result, e.g. that a particular output node produces high values for observations of a particular class. The output weights are learned by means of linear regression and are called readouts since they "read" the reservoir state. Details about the RN training and the echo state property can be found in [9].
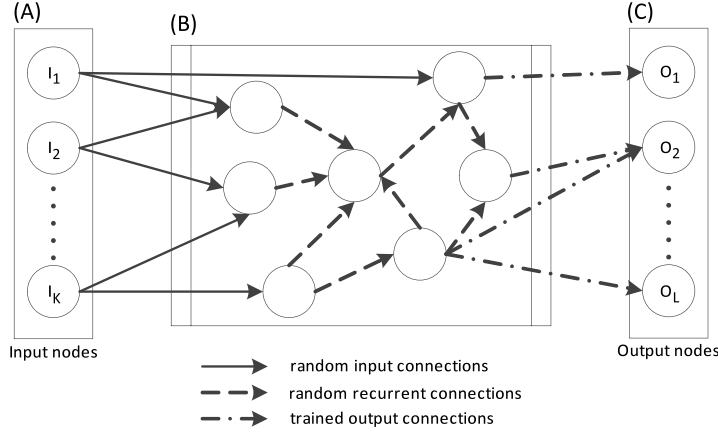
FIGURE 22.3

A standard reservoir network consisting of three layers: (A) the input, (B) the reservoir, and (C) the readout. The second layer includes neurons with non-linear activation functions. The weights of the input and the recurrent connections are randomly fixed. The weights to the output nodes are the only ones being trained.

As a general formulation of the RNs, we assume that the network has K inputs, G neurons (usually called reservoir size), M outputs while the matrices $W_{in}(K \times G)$, $W_{res}(G \times G)$, and $W_{out}(G \times L)$ include the connection weights. The RN system equations are the following:

$$x(k) = f_{res}(W_{in}u(k-1) + W_{res}x(k-1)) \tag{22.1}$$

$$y(k) = f_{out}(W_{out})x(k), \tag{22.2}$$

where $u(k)$, $x(k)$, and $y(k)$ denote the values of the inputs, reservoir outputs, and the readout nodes at time k respectively; f_{res} and f_{out} are the activation functions of the reservoir and the output nodes, respectively. In this work we consider $f_{res}(x) = \tanh(x)$ and $f_{out}(x) = x$ and we fix $L = 5$ equal to the number of the sound classes.

Linear regression is used to determine the weights W_{out},

$$W_{out} = \arg\min_{W}\left(\frac{1}{N_{tr}}\|XW - D\|^2 + \epsilon\|W\|^2\right) \tag{22.3}$$

$$W_{out} = (X^T X + \epsilon I)^{-1}(X^T D), \tag{22.4}$$

where XW and D are the computed vectors, I a unity matrix, N_{tr} the number of the training samples, while ϵ is a regularization term.

The recurrent weights are randomly generated by a zero-mean Gaussian distribution with variance v, which essentially controls the spectral radius (SR) of the reservoir. The largest absolute eigenvalue of W_{res} is proportional to v and is particularly important for the dynamical behavior of the reservoir [8,27]. W_{in} is randomly drawn from a uniform distribution [$-InputScalingFactor$, $+InputScalingFactor$], which emphasizes/deemphasizes the inputs in the activation of the reservoir neurons. The significance of the specific parameter is decreased as the reservoir size increases.

In this work the RN is used for assigning classes to a certain sequence of features coming from audio signals. To this end it is trained so as to achieve an output state where a particular output node is high for observations of a specific class (e.g. Bowhead, Southern Right, etc.) and low for observations of any other class. Thus, the regression layer minimizes the mean squared error between y_t (readout vector) and d_t (desired output), where all the elements belonging to d_t are -1, except the one corresponding to the desired state which is equal to $+1$. Following the work presented in Richard and Lippmann [22], the readout layer $y'_{t,q} = 0.5 + 0.5y_{t,q}$ well approximates the posterior probability vector $P(q|u_t)$, where q corresponds to any given data class.

At this point a small adjustment was introduced ensuring that the probabilities are positive. The readouts are calculated using the following formula:

$$y'_{t,q} = \max(\frac{y_{t,q}+1}{2}, \delta), \ 0 < \delta << 1 \tag{22.5}$$

and the probability is

$$P(u_t|q_t) = \frac{P(q_t, u_t)}{P(q_t)}P(u_t) = \frac{y'_{t,q_t}}{P(q_t)}P(u_t), \tag{22.6}$$

where $P(q_t)$ comprises the mean of $P(q_t|u_t)$ over the training data associated with a specific class. Finally the class associated with the highest $P(q|u_t)$ is assigned to the novel audio signal.

22.5 EXPERIMENTAL SETUP AND RESULTS

This section details the data set used for the experiments, the parameterization of the feature extraction and pattern recognition mechanisms, as well as classification results of the systems considered in this work.

22.5.1 THE DATA SET

We employed the database of the Cooperative Institute for Marine Resources Studies (Oregon State University and NOAA/PMEL) which is publicly available at http://www.mobysound.org/. The following five species were considered: *Blue whales*, *Bowhead whales*, *Fin whales*, *Humpback whales*, and *Southern Right whales*. The recordings are annotated using Region of Interest boxes on the spectrograms of the audio signals. Inside the recordings there may be present more than one call while the annotation files provide the start and end times along with lowest and highest frequency of the call. The sound files are provided in WAVE (.wav) or AIFF (.aif) format. Unlike Halkias et al. [5], we employed the entire data set, even though it is highly unbalanced (see Table 22.2) since our methodology may address this issue. Furthermore, Table 22.2 includes the average SNR of each class of the data set, the difference of which complicates the automatic categorization of the signals. It should be mentioned that we chopped the recordings according to their respective labels in time and used them without artificially adding any noise source. This is motivated by the fact that sound recognition assumes signals coming from a specific class follow a consistent *pattern*, which is learned when both feature extraction and modeling phases operate on the structure of the sound event alone. In addition the recordings were

Table 22.2 The Characteristics of the Data Set of the Cooperative Institute for Marine Resources Studies (Oregon State University and NOAA/PMEL) Which Is Publicly Available at http://www.mobysound.org/

Mysticete Class	SNR (dB)	Duration (s)
Blue whale	8.4	22680
Bowhead whale	16.3	5040
Fin whale	18.7	9720
Humpback whale	27.9	3300
Southern Right whale	12	78

normalized for removing any possible DC-offset. It should be noted that the tenfold cross validation partition scheme was employed and kept the same during the entire experimental campaign.

Our methodology was applied to two tasks: (a) on specific frequency range, and (b) on all five species following the experimental setup presented in [5]. The second task is only an exemplary scenario of the capabilities of the proposed framework since it is impossible for the same microphone to capture sounds coming from all these species at the same habitat.

Some of the recordings may be of poor quality, however they could be useful for creating an automatic call recognition system since training data similar to the ones captured in real-life are required. The data set includes long periods of time where no calls of interest occur, a case which is usually encountered as the events of interest are rather scattered. Furthermore, each recording may include calls from one or more individuals in the present of noise and/or interfering sounds, a characteristic which makes the problematic quite challenging and allows us to thoroughly examine the capabilities of the proposed methodology.

22.5.2 FRAMEWORK PARAMETERIZATION

The feature extraction methods operated on a frame of 30 ms with 10 ms overlap, thus eliminating the existence of possible misalignments. The FFT size, where applicable, was 512 while the testing protocol was the 30-fold cross-validation. Both the training and the testing sets were kept constant between different framework configurations in order to derive comparable results.

The parameters of the RN were selected by means of exhaustive search. They were taken from the following sets: $SR \in \{0.8, 0.9, 0.95, 0.99\}$, $L \in \{100, 500, 1000, 5000, 10000\}$, and *InputScalingFactor* $\in \{0.1, 0.5, 0.7, 0.95, 0.99\}$. The implementation of the RN was based on the echo state network toolbox which is available at http://reservoir-computing.org/software. Model training and parameterization are the most computationally intensive parts of the methodology proposed here; however, they are performed only once and off-line.

22.5.3 EXPERIMENTAL RESULTS AND ANALYSIS

The experimental campaign is comprised of two phases: (a) the first one includes all the five mysticete species, and (b) the second one concerns the classification of species with acoustic content ranging within specific frequency boundaries, thus a smaller amount of classes is taken into account. It should be noted that the first phase comprises a hypothetical scenario since is highly unlikely that the selected

Table 22.3 The Average Recognition Rates (in %) With Respect to Every Feature Set, While the Highest One Is Emboldened. Each Experiment Was Repeated 50 Times While the Considered Species Are the Following: *Bowhead*, *Humpback*, *Southern Right*, *Blue*, and *Fin*

Classifier	Feature Set		
	Frequency	**Wavelet**	**Concatenation**
Reservoir network	78.1 ± 0.9	73.2 ± 0.8	$\mathbf{81.3 \pm 0.8}$

Table 22.4 The Confusion Matrix Which Includes the Classification Rates Reached by the RN Modeling the Concatenated Feature Set

Presented	Responded				
	Bowhead	**Humpback**	**S. Right**	**Blue**	**Fin**
Bowhead	$\mathbf{76.2 \pm 0.5}$	8.1 ± 0.3	7 ± 0.2	7.2 ± 1.9	1.5 ± 0.6
Humpback	12.1 ± 1.5	$\mathbf{70.8 \pm 0.3}$	9.4 ± 0.8	7.7 ± 1.2	0 ± 0.7
S. Right	7.6 ± 0.3	7.7 ± 0.8	$\mathbf{79.8 \pm 0.9}$	0 ± 0.4	4.5 ± 0.6
Blue	10.2 ± 0.3	0 ± 1.5	7.5 ± 0.4	$\mathbf{82.3 \pm 0.5}$	0 ± 1.4
Fin	8.3 ± 0.4	4 ± 0.2	0 ± 0.8	6.7 ± 0.7	$\mathbf{81 \pm 1.9}$

five species will coexist in the same habitat; however, it may be useful to provide indications regarding the efficacy of the acoustic characterization of whale vocalizations.

The concentrated results of the first phase are tabulated in Table 22.3. It includes the average recognition rates with respect to each feature set, while it considers all the species (*Bowhead*, *Humpback*, *Southern Right*, *Blue*, and *Fin*). The confidence intervals computed over 50 iterations per experiment are also shown. As we can see, the concatenated feature set provides better performance. Furthermore, the set derived from the frequency domain achieves superior recognition rates with respect to the one based on the wavelet domain. The performance of the system is placed at quite satisfactory levels (average rate is equal to 78%) due to the simultaneous usage of sound parameters coming from multiple domains. This fact suggests that the feature sets exhibit complementary characteristics which are beneficial and, in the end, lead to higher classification rates.

Interestingly the RN offers very good classification accuracy showing that when a vast amount of data is available, the RN technique is able to learn to discriminate the data belonging to different classes, i.e. the boundaries within the high-dimensional feature space.

Moving on, we employed confusion matrices which may provide a better understanding of the system learning capacities (Tables 22.4, 22.5, 22.6). Following the results of the first experimental phase, we employed the RN operating on the concatenated feature space for the purposed of the second one. The involved tasks are two: (a) classification between the *Southern Right*, *Humpback*, *Bowhead* species (Table 22.6), and (b) classification between the *Blue* and *Fin* species (Table 22.5). The proposed approach reaches quite high recognition rates: 78.7% for the first task and 91.8% for the second one. The discrimination between Blue and Fin whales is more than satisfactory while the species with the lowest rate is the Humpback whale with $72 \pm 0.9\%$ since it is misinterpreted for the Bowhead species. The species with the highest recognition accuracy is the Blue whale and the one with the lowest the Humpback.

Overall we infer that the proposed system achieves quite high recognition rates considering the acoustic similarities between the repertoire of different or even the same species. This fact is actually

Table 22.5 The Confusion Matrix Which Includes the Classification Rates Reached by the RN Modeling the Concatenated Feature Set While the Considered Species Are *Blue* and *Fin Whale*

Presented	Responded	
	Blue	**Fin**
Blue	**92 ± 0.3**	8 ± 0.5
Fin	8.5 ± 0.2	**91.5 ± 0.4**

Table 22.6 The Confusion Matrix Which Includes the Classification Rates Reached by the RF Fusion Modeling the Concatenated Feature Set While the Considered Species Are *Southern Right*, *Humpback*, and *Bowhead Whale*

Presented	Responded		
	Southern Right	**Humpback**	**Bowhead**
Southern Right	**80 ± 1**	13.2 ± 0.6	6.8 ± 0.2
Humpback	12.1 ± 0.4	**72 ± 0.9**	15.9 ± 0.4
Bowhead	8.3 ± 0.2	7.6 ± 0.3	**84.1 ± 0.7**

a challenge also for trained human annotators. A direct comparison with the work reported in [5] is not feasible since both training sets and regimes are different. The present framework captures diverse characteristics of the audio content combined with a powerful pattern recognition algorithm, which are essential for a reliable discrimination.

22.6 CONCLUSIONS

This article analyzed a novel methodology for the automatic classification of Mysticete sounds. It has proven to be able to provide encouraging classification rates in multiple tasks including 2, 3, and 5 classes. The system can handle more species given an adequate amount of related training data. It may comprise a useful tool toward research conducted on as well as off the field. An interesting point is that even though the RN is basically untrained (only the readout layer is trained by means of linear regression) it performs in a quite satisfactory manner.

The article provided a good proof of concept of the applicability of the audio recognition technology onto the specific problem, which may encourage further research in this thematic. Future work includes (a) the development of a statistical noise elimination mechanism, (b) the design of a preprocessing stage for discriminating between known and unknown (previously seen and not seen) data for deciding whether the specific data sequence can be processed by the system or it is part of a completely new class (novelty detector [16]), and (c) deriving a metric measuring the distance between the unknown data and forming new classes (which may be new species or previously unseen parts of the repertoires of known species) since it is unrealistic to assume that we have gathered data representing every manifestation of the Mysticete species.

ACKNOWLEDGMENTS

REFERENCES

[1] M. Bahoura, Y. Simard, Blue whale calls classification using short-time Fourier and wavelet packet transforms and artificial neural network, Digit. Signal Process. 20 (4) (2010) 1256–1263.

[2] S. Baluja, M. Covell, Waveprint: efficient wavelet-based audio fingerprinting, Pattern Recognit. 41 (11) (2008) 3467–3480.

[3] J.C. Brown, P.J.O. Miller, Automatic classification of killer whale vocalizations using dynamic time warping, J. Acoust. Soc. Am. 122 (2) (2007) 1201–1207.

[4] J.C. Brown, P. Smaragdis, Hidden Markov and Gaussian mixture models for automatic call classification, J. Acoust. Soc. Am. 125 (6) (2009).

[5] X.C. Halkias, S. Paris, H. Glotin, Classification of mysticete sounds using machine learning techniques, J. Acoust. Soc. Am. 134 (5) (2013).

[6] F. Harris, On the use of windows for harmonic analysis with the discrete Fourier transform, Proc. IEEE 66 (1) (1978) 51–83.

[7] S.B. Holmes, K.A. McIlwrick, L.A. Venier, Using automated sound recording and analysis to detect bird species-at-risk in southwestern Ontario woodlands, Wildl. Soc. Bull. (2014), http://dx.doi.org/10.1002/wsb.421.

[8] H. Jaeger, Tutorial on Training Recurrent Neural Networks, Covering BPPT, RTRL, EKF and the "Echo State Network" Approach, Tech. rep., Fraunhofer Institute AIS, St. Augustin, Germany, 2002.

[9] M. Lukoševičius, H. Jaeger, Reservoir computing approaches to recurrent neural network training, Comput. Sci. Rev. 3 (3) (2009) 127–149.

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

[11] D.K. Mellinger, C.W. Clark, Recognizing transient low-frequency whale sounds by spectrogram correlation, J. Acoust. Soc. Am. 107 (6) (2000) 3518–3529.

[12] X. Mouy, M. Bahoura, Y. Simard, Automatic recognition of fin and blue whale calls for real-time monitoring in the St. Lawrence, J. Acoust. Soc. Am. 126 (6) (2009).

[13] S.O. Murray, E. Mercado, H.L. Roitblat, The neural network classification of false killer whale (Pseudorca crassidens) vocalizations, J. Acoust. Soc. Am. 104 (6) (1998).

[14] S. Ntalampiras, N. Fakotakis, Speech/music discrimination based on discrete wavelet transform, in: Proceedings of the 5th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications, SETN '08, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 205–211.

[15] S. Ntalampiras, I. Potamitis, N. Fakotakis, Acoustic detection of human activities in natural environments, J. Audio Eng. Soc. 60 (9) (2012) 686–695, http://www.aes.org/e-lib/browse.cfm?elib=16373.

[16] M.A. Pimentel, D.A. Clifton, L. Clifton, L. Tarassenko, A review of novelty detection, Signal Process. 99 (0) (2014) 215–249, http://www.sciencedirect.com/science/article/pii/S016516841300515X.

[17] I. Potamitis, T. Ganchev, N. Fakotakis, Automatic acoustic identification of crickets and cicadas, in: 9th International Symposium on Signal Processing and Its Applications, ISSPA 2007, Feb. 2007, pp. 1–4.

[18] I. Potamitis, S. Ntalampiras, O. Jahn, K. Riede, Automatic bird sound detection in long real-field recordings: applications and tools, Appl. Acoust. 80 (2014) 1–9.

[19] I. Potamitis, I. Rigakis, Large aperture optoelectronic devices to record and time-stamp insects wingbeats, IEEE Sens. J. 16 (15) (2016) 6053–6061.

[20] L. Ranjard, S.J. Withers, D.H. Brunton, H.A. Ross, S. Parsons, Integration over song classification replicates: song variant analysis in the hihi, J. Acoust. Soc. Am. 137 (5) (2015).

[21] Y. Ren, M.T. Johnson, J. Tao, Perceptually motivated wavelet packet transform for bioacoustic signal enhancement, J. Acoust. Soc. Am. 124 (1) (2008).

[22] M. Richard, R. Lippmann, Neural net classifiers estimate posterior probabilities, Neural Comput. 3 (4) (1991) 461–483.

[23] L. Shamir, C. Yerby, R. Simpson, A.M. von Benda-Beckmann, P. Tyack, F. Samarra, P. Miller, J. Wallin, Classification of large acoustic datasets using machine learning and crowdsourcing: application to whale calls, J. Acoust. Soc. Am. 135 (2) (2014) 953–962.

[24] S. Shamma, On the role of space and time in auditory processing, Trends Cogn. Sci. 5 (8) (2001) 340–348.

[25] S.V. David, N. Mesgarani, S.A. Shamma, Estimating sparse spectro-temporal receptive fields with natural stimuli, Netw. Comput. Neural Syst. 18 (3) (2007) 191–212.

[26] F. Triefenbach, A. Jalalvand, K. Demuynck, J.-P. Martens, Acoustic modeling with hierarchical reservoirs, IEEE Trans. Audio Speech Lang. Process. 21 (11) (2013) 2439–2450.

[27] D. Verstraeten, B. Schrauwen, M. D'Haene, D. Stroobandt, An experimental unification of reservoir computing methods, Neural Netw. 20 (3) (2007) 391–403.

[28] W.S.D. Wilcock, Tracking fin whales in the Northeast Pacific Ocean with a seafloor seismic network, J. Acoust. Soc. Am. 132 (4) (2012).