A Deep Learning Framework for Classifying Mysticete Sounds

Stavros Ntalampiras

Politecnico di Milano
Department of Electronics, Information and Bioengineering
Piazza Leonardo da Vinci, 32 I-20133, Milan, Italy
e-mail: stavros.ntalampiras@polimi.it, dalaouzos@gmail.com
site: `sites.google.com/site/stavrosntalampiras/home`

Abstract

This paper addresses a problem belonging to the domain of whale audio processing, more specifically the automatic classification of sounds produced by the Mysticete species. The specific task is quite challenging given the vast repertoire of the involved species, the adverse acoustic conditions and the nearly inexistent prior scientific work. Two feature sets coming from different domains (frequency and wavelet) were designed to tackle the problem. These are modeled by an Echo State Network classifier. The dataset includes five species (*Blue*, *Fin*, *Bowhead*, *Southern Right*, and *Humpback*) and it is publicly available at `http://www.mobysound.org/`. We followed a thorough experimental procedure and achieved more than satisfactory recognition rates.

Keywords: marine mammal acoustic signal processing, multidomain parameters, reservoir network, deep learning.

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, e.g. (Holmes et al., 2014; Potamitis et al., 2007; Potamitis and Rigakis, 2016). The ultimate goal of frameworks processing bioacoustic signals is to provide a complete and accurate picture of the biodiversity of the habitat of interest towards its conservation (Ntalampiras et al., 2012). 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 (Potamitis et al., 2014; Ranjard et al., 2015). 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 towards 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 (Shamma, 2001; Stephen V. David and Shamma, 2007). This results on a consistent distribution of the energy produced by each audio signal with respect to specific parts of the spectrum. Our aim 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 co-existing 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*) dataset 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 dataset and reached quite encouraging classification rates.

The rest of this article is organized as follows: Section 2 provides an overview of the related literature. Section 4 analyses 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 6 offers our conclusions as well as ideas for future works.

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. (Halkias et al., 2013) 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 Boltzman 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 (Mouy et al., 2009). The methods include spectrogram matching, dynamic time wrapping and vector quantization while the latter two operate on the frequency contour. The dataset was recorded by the authors while they emphasize on the strong and weak points of each method.

Study (Shamir et al., 2014) 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 (Bahoura and Simard, 2010) the authors used the short-time Fourier and wavelet packet transforms along with a multilayer 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 (Murray et al., 1998) employs two types of neural networks (based on competitive learning and Kohonen feature mapping respectively) in order to analyse 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 et al. (Brown and Miller, 2007) 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 (Brown and Smaragdis, 2009) 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 (Mellinger and Clark, 2000) based on spectrogram correlation. The corpus consisted of bowhead whale (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. (M. A. Roch and Hildebrand, 2007) 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 (Wilcock, 2012) followed a fundamentally different approach and performed tracking of fin whales in the northeast 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.

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 towards 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. 1.

3.1. Frequency domain features

The first feature set exploits the spectrogram of the sound event since is 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 (Harris, 1978). Here we have employed the following windowing techniques to reduce edge effects in the FFT: *a*) blackman, *b*) hamming, *c*) hanning, *d*) and 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 windowing offers the best spectral representation with respect to classification accuracy, thus it is favored in this work. As one may see in Fig. 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 commonly used in audio processing applications). Thus the final feature vector is the energies of the short time Fourier transform after it is Blackman windowed.

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 (Ren et al., 2008), audio fingerprinting (Baluja and Covell, 2008), speech/music discrimination (Ntalampiras and Fakotakis, 2008) etc.). The uniqueness of

the wavelet transform come 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. 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 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 ($5 \times 8 = 40$). This is the WP-integration feature vector and the block diagram for its computation is demonstrated in Fig. 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.

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:

4.1. Reservoir network

The trend in acoustic modeling suggests the usage of Reservoir Computing (RC) techniques (Triefenbach et al., 2013). 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. 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 read-outs since they "read" the

Table 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

reservoir state. Details about the RN training and the echo state property can be found in (Lukoeviius and Jaeger, 2009).

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:

$$x(k) = f_{res}(W_{in}u(k-1) + W_{res}x(k-1)) \quad (1)$$
$$y(k) = f_{out}(W_{out})x(k), \quad (2)$$

where $u(k)$, $x(k)$ and $y(k)$ denote the values of the inputs, reservoir outputs and the read-out 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},

$$W_{out} = \underset{W}{argmin}(\frac{1}{N_{tr}}\|XW - D\|^2 + \epsilon\|W\|^2) \quad (3)$$
$$W_{out} = (X^T X + \epsilon I)^{-1}(X^T D), \quad (4)$$

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 (Jaeger, 2002; Verstraeten et al., 2007). W_{in} is randomly drawn from a uniform distribution $[-InputScalingFactor, +InputScalingFactor]$, which emphasises/deemphasises 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. hungry, sleepy, etc.) and low for observations of any other class. Thus, the regression layer minimizes the mean squared error between y_t (read-out 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, 1991) the read-out 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 read-outs are calculated using the following formula:

$$y'_{t,q} = \max(\frac{y_{t,q} + 1}{2}, \delta), 0 < \delta << 1 \tag{5}$$

and the probability is

$$P(u_t|q_t) = \frac{P(q_t, u_t)}{P(q_t)} P(u_t) = \frac{y'_{t,q_t}}{P(q_t)} P(u_t), \tag{6}$$

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.

5. Experimental Set-up and Results

This section details the dataset 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.

5.1. The dataset

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., 2013) we employed the entire dataset, even though it is highly unbalanced (see Table 2) since our methodology may address this issue. Furthermore Table 2 includes the average SNR of each class of the dataset, the

Table 2: The characteristics of the dataset 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

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 normalized for removing any possible DC-offset. It should be noted that the ten-fold 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 set-up presented in (Halkias et al., 2013). 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 dataset 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.

Table 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	**81.3 ± 0.8**

Table 4: The confusion matrix which includes the classification rates reached by the RN modelling the concatenated feature set.

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

5.2. Framework parametrization

The feature extraction methods operated on a frame of 30ms with 10ms overlap, thus a 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 parametrization are the most computationally intensive parts of the methodology proposed here, however they are performed only once and off-line.

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 clas-

sification 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 five species will co-exist 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 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. 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 6), and *b*) classification between the *Blue* and *Fin* species (Table 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 a challenge also for trained human annotators. A direct comparison with the work reported in (Halkias

et al., 2013) 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.

6. Conclusions

This article analysed 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 towards 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 problematic, 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 (Pimentel et al., 2014)), and *c*) derive 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.

Table 5: The confusion matrix which includes the classification rates reached by the RN modelling the concatenated feature set while the considered species are *Blue* and *Fin whale*.

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

Acknowledgment

The author would like to dedicate this work to Mrs. Marigianna Kotta of the IOW Leibniz Institute for Baltic Sea Research (`http://www.io-warnemuende.de/en_index.html`).

In addition, the author is thankful to Sarah Heimlich, Holger Klinck and Dave Mellinger for `http://www.mobysound.org/` which provided the dataset for conducting the experiments included in this work.

References

Bahoura, M., Simard, Y., 2010. Blue whale calls classification using short-time fourier and wavelet packet transforms and artificial neural network. Digital Signal Processing 20 (4), 1256 – 1263.

Baluja, S., Covell, M., 2008. Waveprint: Efficient wavelet-based audio fingerprinting. Pattern Recognition 41 (11), 3467 – 3480.

Brown, J. C., Miller, P. J. O., 2007. Automatic classification of killer whale vocalizations using dynamic time warping. The Journal of the Acoustical Society of America 122 (2), 1201–1207.

Brown, J. C., Smaragdis, P., 2009. Hidden markov and gaussian mixture models for automatic call classification. The Journal of the Acoustical Society of America 125 (6).

Halkias, X. C., Paris, S., Glotin, H., 2013. Classification of mysticete sounds using machine learning techniques. The Journal of the Acoustical Society of America 134 (5).

Table 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*.

Responded / **Presented**	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**

Harris, F., Jan 1978. On the use of windows for harmonic analysis with the discrete fourier transform. Proceedings of the IEEE 66 (1), 51–83.

Holmes, S. B., McIlwrick, K. A., Venier, L. A., 2014. Using automated sound recording and analysis to detect bird species-at-risk in southwestern ontario woodlands. Wildlife Society Bulletin, n/a–n/a.

Jaeger, H., 2002. Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the echo state network approach. Tech. rep., Fraunhofer Institute AIS, St. Augustin-Germany.

Lukoeviius, M., Jaeger, H., 2009. Reservoir computing approaches to recurrent neural network training. Computer Science Review 3 (3), 127 – 149.

M. A. Roch, M. S. Soldevilla, J. C. B. E. H., Hildebrand, J. A., 2007. Gaussian mixture model classification of odontocetes in the southern california bight and the gulf of california. The Journal of the Acoustical Society of America 121 (3), 1737–1748.

Mellinger, D. K., Clark, C. W., 2000. Recognizing transient low-frequency whale sounds by spectrogram correlation. The Journal of the Acoustical Society of America 107 (6), 3518–3529.

Mouy, X., Bahoura, M., Simard, Y., 2009. Automatic recognition of fin and blue whale calls for real-time monitoring in the st. lawrence. The Journal of the Acoustical Society of America 126 (6).

Murray, S. O., Mercado, E., Roitblat, H. L., 1998. The neural network classification of false killer whale (pseudorca crassidens) vocalizations. The Journal of the Acoustical Society of America 104 (6).

Ntalampiras, S., Fakotakis, N., 2008. 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, pp. 205–211.
URL http://dx.doi.org/10.1007/978-3-540-87881-0_19

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

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

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

Potamitis, I., Ntalampiras, S., Jahn, O., Riede, K., 2014. Automatic bird sound detection in long real-field recordings: Applications and tools. Applied Acoustics 80 (0), 1 – 9.

Potamitis, I., Rigakis, I., Aug 2016. Large aperture optoelectronic devices to record and time-stamp insects wingbeats. IEEE Sensors Journal 16 (15), 6053–6061.

Ranjard, L., Withers, S. J., Brunton, D. H., Ross, H. A., Parsons, S., 2015. Integration over song classification replicates: Song variant analysis in the hihi. The Journal of the Acoustical Society of America 137 (5).

Ren, Y., Johnson, M. T., Tao, J., 2008. Perceptually motivated wavelet packet transform for bioacoustic signal enhancement. The Journal of the Acoustical Society of America 124 (1).

Richard, M., Lippmann, R., 1991. Neural net classifiers estimate posterior probabilities. Neural Computation 3 (4), 461483.

Shamir, L., Yerby, C., Simpson, R., von Benda-Beckmann, A. M., Tyack, P., Samarra, F., Miller, P., Wallin, J., 2014. Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. The Journal of the Acoustical Society of America 135 (2), 953–962.

Shamma, S., 2001. On the role of space and time in auditory processing. Trends in Cognitive Sciences 5 (8), 340–348.

Stephen V. David, N. M., Shamma, S. A., 2007. Estimating sparse spectro-temporal receptive fields with natural stimuli. Network: Computation in Neural Systems 18 (3), 191–212.

Triefenbach, F., Jalalvand, A., Demuynck, K., Martens, J.-P., Nov 2013. Acoustic modeling with hierarchical reservoirs. Audio, Speech, and Language Processing, IEEE Transactions on 21 (11), 2439–2450.

Verstraeten, D., Schrauwen, B., DHaene, M., Stroobandt, D., 2007. An experimental unification of reservoir computing methods. Neural Networks 20 (3), 391 – 403, ¡ce:title¿Echo State Networks and Liquid State Machines¡/ce:title¿.

Wilcock, W. S. D., 2012. Tracking fin whales in the northeast pacific ocean with a seafloor seismic network. The Journal of the Acoustical Society of America 132 (4).

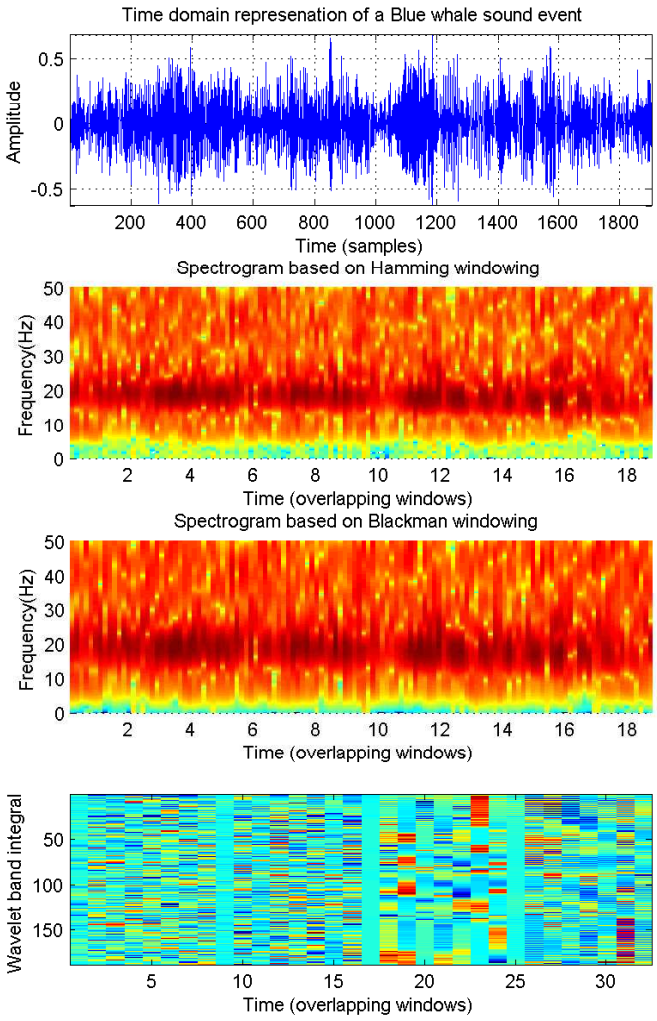

Figure 1: A representation of the feature sets used in this work. Both frequency and wavelet domains are considered.

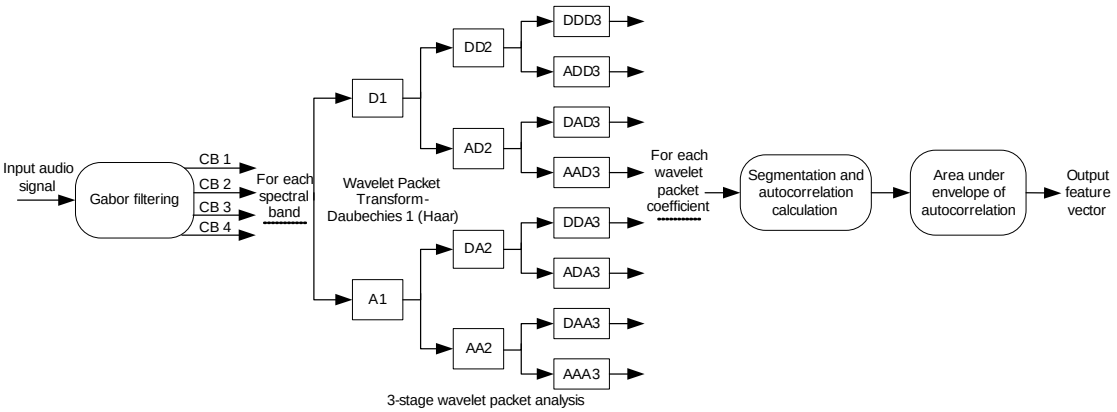

Figure 2: The block diagram of the process extracting the wavelet packet integration feature set.

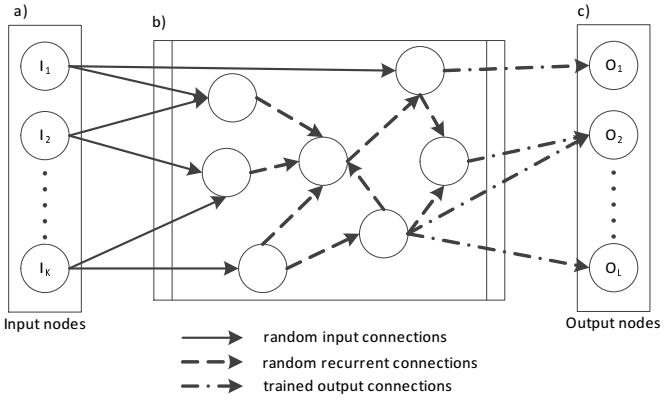

Figure 3: A standard reservoir network consisting of three layers: a) the input, b) the reservoir and c) the readout. The second layer includes neurons with non-linear activation functions. The weights of the input and the recurrent connections are randomly fixed. The weights to the output nodes are the only ones being trained.