A New Approach for North Atlantic Right Whale Upcall Detection

Ali K Ibrahim, Hanqi Zhuang, Nurgun Erdol, and Ali Muhamed Ali
Department Computer & Electrical Engineering and Computer Science
Florida Atlantic University
Boca Raton, FL 33431
{aibrahim2014, zhuang, erdol, amuhamedali2014}@fau.edu

***Abstract*— In this paper, a new up-call detection algorithm for North Atlantic Right Whales is proposed. Mel Frequency Cepstral Coefficients and Discrete Wavelet Transform (DWTs) are used to extract features of North Atlantic Right Whale up-calls. Several types of wavelets are tested in terms of detection accuracy. Once the up-call features are extracted, classifiers such as Support Vector Machines and KNNs are applied to classify call types. Detection results shows that the upcall detection rate by using MFCCs is 73.82%. However, the upcall detection rate is improved to 92.27% when MFCCs are integrated with DWTs for feature extraction. Furthermore, it is shown through experimental studies that the proposed detection scheme is superior to those obtained using spectrogram-based techniques in terms of both detection accuracy and speed.**

***Keywords*— North Atlantic Right Whale, DWT, MFCC, Upcall Detection, DSP, Pattern Recognition**

I. INTRODUCTION

Marine mammals spend most of their time under water and many of them, including sperm whales, are deep divers so that detection based on their visual signals proves to be difficult, unreliable and very expensive. They are, however, very vocal and use acoustic signals to communicate and socialize with each other and to navigate making Passive Acoustic Monitoring (PAM) [1, 2] the technology of choice for their detection and tracking. Passive acoustic monitoring is conducted by placing static hydrophone buoys in areas of interest to monitor acoustic activity. Combined with automatic detection software the buoys can log the presence of different marine mammal species and are an important tool in studying cetacean distribution and abundance. To facilitate measures that protect and ensure the safety of these animals it is important to determine if high-risk areas encroach into their migration paths. Such measures would prevent, for example, accidental encounters with ships and fishing nets, or protect them from anthropogenic causes that lower their rates of reproduction and lead them to extinction.

The North Atlantic Right Whale (NARW) is one of the critically endangered whales whose low birth rate cannot compensate its death rate [1]. The number of NARWs recently recorded in the east coast of North America is only about 300-500 [3]. Tracking their numbers and migration paths in order to prevent accidental deaths and promote their reproduction is vital to sustaining their existence. Acoustic monitoring of NARWs is accomplished by detecting their signature vocalizations, named upcalls, which are narrow band signals with frequency swings in the range of 50-250 Hz [4]. Early attempts at detection consisted of single-stage algorithms that used edge detection [5] and time-frequency convolution [6]. They had relatively high levels of false positive errors [7]. Later methods with feature extraction and classification [7, 8] capabilities were able to reduce the number of false positives although the numbers remained high enough to be considered significant [7]. In our earlier work [9, 10], we treated NARW vocalization spectrograms as images and extracted space-time features for classification. We tested time-frequency parameters extracted from upcall spectrogram contours, and texture data output of a Linear Binary Pattern (LBP) operator acting on the spectrogram. We reported detection accuracies of 93-99% with a combination of the two feature sets and the K-nearest neighbor and Support Vector Machine classification algorithms. Our study established the significance of a match between features and classification algorithms. The use of image processing tools to extract features from NARW vocalization spectrograms is very intuitive and inspired by the ability of a human operator to do the same by a simple observation the spectrograms. The method achieves high levels of detection accuracy but requires $O(\frac{N}{1-p})$ data points for a call recording of length N arranged in frames with overlap fraction $0.5<p<1$. A typical call is 2 seconds long and is sampled at 2KHz resulting in a spectrogram with more than 8000 samples is large and difficult to manipulate.

In this work, the problem of complexity due to data size is circumvented by working directly on the signal in frames of size 250 ms. or 500 samples. Inspired by the success of Mel Frequency Cepstral Coefficients (MFCC) in speech recognition, but cognizant of their sensitivity to noise, we propose features consisting of MFCCs computed in wavelet subspaces. Classification by Support Vector Machines (SVMs) follows.

There is no particular reason that the Cepstral coefficients should be derived from the mel frequency spectra as opposed to any log spectra. Mel frequencies are approximately logarithmic and are used here for convenience. Wavelet subspaces are introduced for purposes of de-

978-1-5090-3071-2/16 $31.00 © 2016 IEEE
DOI 10.1109/IS3C.2016.76

noising the upcalls from ambient ocean noise, a process that ensures the robustness of the MFCCs. The structures of the wavelet filters and kernel transformations that enable hyperplane decision boundaries are investigated to improve the effectiveness NARW upcall detection.

II. Feature Extraction

The NARW detection algorithm consists of three typical steps shown in Fig.1. Preprocessing limits the signal frequency range to a cautious upcall range of 20-450 Hz. The feature extraction stage consists of a Discrete Wavelet Transform followed by the computation of the Mel Frequency Cepstral Coefficients (MFCCs). The Support Vector Machine (SVM) classifier determines if the recorded audio signal contains an NARW upcall.

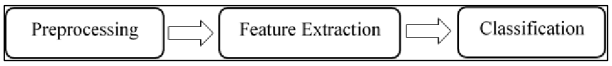

Fig. 1. The proposed NARW upcall detection scheme

MFCCs are features widely used in automatic speech recognition systems. Mel scale consists of pitch frequencies judged by human listeners to be equidistant from one another. They are also commonly used to identify vocalizations of marine mammals [11]. MFCC values are not robust in the presence of noise. Due to very low signal-to-noise-ratios encountered in the ocean environment, any effort to use MFCCs as features for recognition must be accompanied with noise reduction. It is well known that in many instances of signal and noise constructs, restriction of the noisy signal to a particular subspace may significantly increase the SNR especially if the subspace is characterized by a concentration of the signal but not of the noise.

A typical NARW upcall spectrogram is shown in Fig. 2. The noise spectral estimate derived from the same data is shown in Fig. 3. It shows that noise spectral density is nearly constant over the frequency range not affected by the data acquisition system's antialiasing filter. To maximize the local SNR, we made use of wavelet filter banks [12-15]. We kept only the approximation coefficients of an $N \in [2,3,4]$ stage DWT. We subsequently computed the MFCCs of the approximation coefficients and used the MFCCs as feature vectors for classification. The approximation coefficients of the transform are the output of the wavelet scaling filter of impulse response $g[n]$ shown in Fig. 4. For best SNR results, $g[n]$ must be matched to the signal. Rather than taking a model based approach, we experimented with a selection of waveletes to find the best among them. The order of the DWT determines the bandwidth or scale of the filter and is expected to have an effect on the output SNR as well as the quality of the MFCC for correct classification.

The algorithm that takes the noisy signal $x[n]$ (vocalization signal $s[n]$ corrupted with additive noise $d[n]$) and extracts features for classification can be listed as follows:

1. Divide $x[n]$ into 250 millisecond frames $x_i[n]$ overlapping by 50%.
2. Decompose each frame using an N-stage DWT. Keep only the approximation coefficients $a_i[n]$.
3. Compute the DFT of $a_i[n] \leftrightarrow A_i[k]$, for n and $k \in [0, L-1]$.
4. Let $B_i[k] = |A_i[k]|^2 \leftrightarrow b_i[n]$. Note that $\{b_i[n]\}_n$ is the autocorrelation sequence of the coefficients $a_i[n]$.
5. Compute the output $y_{il} = \frac{1}{L}\sum_{k=0}^{L-1} B_i[k]V_l[k]$ of Mel filters $V_l[k] = V_0\left[k - \frac{\omega_l}{2\pi}\right]$. This is the Mel spectrum which, roughly speaking, is the spectrum of the approximation coefficients at Mel frequencies.
6. The MFCCs are the DCT of the log Mel spectrum.

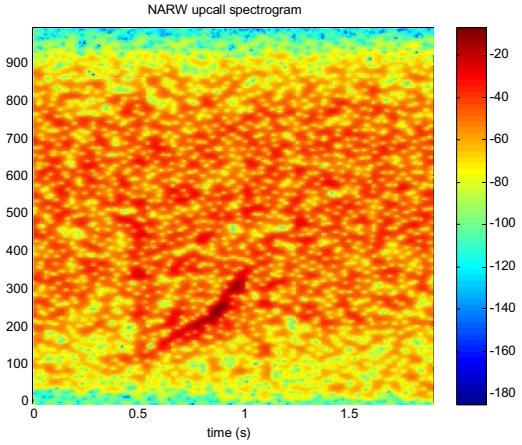

Fig. 2. A typical NARW spectrogram showing the upcall range

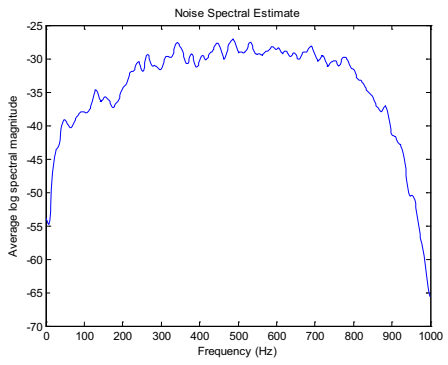

Fig. 3. Spectral estimate of the ocean ambient noise

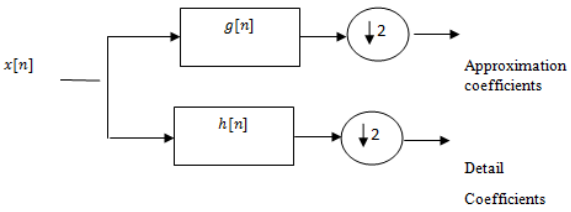

Fig. 4. A single-stage wavelet decomposition filter bank

III. Overview of Support Vector Machine Classifier

In this section, we present a brief overview of the Support Vector Machine (SVM) classifier for readability. Readers are referred to [16-18] for a more detailed description.

SVM, derived from the theory of Structural Risk Minimization, was first introduced by Vapnik [16]. Detailed information and further references can be found in [17, 18]. SVMs are classifiers that can be separate objects into their respective groups using lines or hyperplanes that are derived from the objects. When it is not possible for a straight line to separate the objects, a kernel transformation can rearrange the objects so that their separation by a hyperplane is possible. The problem then becomes that of selecting a proper kernel [16]. Important kernel functions which fulfill these conditions are the polynomial kernel, which maps the data into the space of all polynomials of degree d, and the Gaussian radial basis function (RBF) kernel. Given the training set $\{x_i, y_i\}$ consisting of training vectors $x_i \in R_n$ and their corresponding labels $y_i \in \{-1, +1\}$, a kernel function $K(x, y)$ and a parameter C, the SVM classifier finds an optimal separating hyperplane in F. This is done by solving the following quadratic programming problem: Choosing the vector α which maximizes

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{J=1}^{N} \alpha_i \alpha_j \, y_i y_j k(x_i, x_j)$$

under the constraints

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad and \; 0 \le \alpha_i \le C \quad \forall i.$$

The parameter $C > 0$ allows us to specify how strictly we want the classifier to fit to the training data (a larger C meaning more strictly). The output of the SVM is

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x)$$

where $f(x) > 0$ means that x is classified to class +1. The training vectors x_i for which the α_i are greater than zero are called support vectors.

IV. Detection Results

The effectiveness of the features described in Section II is evaluated for NARW upcall detection with various popular classifiers initially. Three thousands right whale audio segments were used in the training phase, in which there were 241 NARW upcalls and 2759 non-upcalls. In the verification phase, 1500 NARW audio segments were used for the purpose of upcall detection, among which, 233 are considered upcall audio segments, the rest (1268) are non-up calls. The following formulas are used to determine the accuracy of the system:

$$\text{Overal detection rate} = \frac{\text{number of correctly classified calls}}{\text{total number of calls}}$$

$$\text{Upcall detection rate} = \frac{\text{number of correctly classified upcalls}}{\text{total number of upcalls}}$$

$$\text{Non-upcall detection rate} = \frac{\text{number of correctly classified non-upcalls}}{\text{total number of non-upcalls}}$$

It has been found in this research that for classification, SVMs fare better in comparison to other algorithms such as KNN due to the fact that SVMs perform well on datasets that have overlapped attributes. As a result, the overall experimental system was built by integrating DWTs, MFCCs and SVMs.

Table 1 shows the detection results obtained from various types of features and the linear SVM. It is observed that the highest rate of correct upcall detection is achieved by using two stage discrete wavelet transform followed by MFCC computation, which is 91.85% and overall detection rate is 94.47%. On the other hand, the MFCC and SVM combination produces the lowest upcall detection rate.

Table 1 Detection Results using MFCCs and DWTs

Type of feature	upcall detection rate	non upcall detection rate	overall detection
MFCC	73.82%	97.16%	93.53%
DWT sigle stage +MFCC	77.68%	97.32%	94.27%
DWT two stages+MFCC	91.85%	94.95%	94.47%
DWT three stages+MFCC	86.27%	91.96%	91.07%

In Table 2, results from using different types of wavelet families (for instance, Symlets and Daubechies) are shown. Table 2 indicates that Daubechies (db4) is the best type of features for upcall detection, which achieves an accuracy of 92.27%. On the other hand, db1 has the lowest detection rate, which is 86.2661%.

Table 2 Detection Results using Different Wavelets

Type of wavelet	Upcall detection rate	non upcall detection rate	overall detection rate
db1	86.27%	93.69%	92.54%
db2	86.70%	94.24%	93.07%
db3	90.56%	95.03%	94.34%
db4	92.27%	94.87%	94.47%
db5	88.41%	93.69%	92.87%
db6	90.13%	95.35%	94.54%
db7	91.41%	95.43%	94.80%
coif1	88.84%	95.11%	94.14%
coif2	89.70%	94.01%	93.34%
coif3	88.84%	94.72%	93.84%
coif4	90.99%	94.01%	93.54%
coif5	90.13%	95.03%	94.27%
sym2	86.70%	94.24%	93.07%
sym3	90.56%	95.03%	94.34%
sym4	89.27%	94.01%	93.27%
sym5	89.70%	94.24%	95.53%
sym6	86.27%	94.48%	93.20%

Table 3 compares the results obtained in this research project with those reported in [9, 10]. The highest upcall detection rate obtained by using DWTs & MFCCs as features and the linear SVM as the classifier is 92.27%. On the other hand, the best upcall detection rate reported

in [9], which was obtained by using LBP features and linear SVMs, is 90.41%.

There are a few advantages in using the proposed scheme in comparison with that given in [9, 10]. First, in contrast to the 2D spectrogram-based approach designed in [9,10], the features used in this research are extracted from 1D time signals, therefore the computational cost of the algorithms is orders of magnitude lower. Second, the algorithms used in this research produced superior results in terms of both the upcall and overall detection rates.

TABLE 3 DETECTION RESULTS USING DIFFERENT FEATURES AND DIFFERENT CLASSIFIERS

Method	Upcall detection rate	non upcall detection rate	overall detection rate
DWT+MFCC+Linear SVM	92.27%	94.87%	94.47%
TFP-2 features+LDA	80.11%	95.21%	91.70%
TFP-2 features+QDA	67.38%	92.78%	86.87%
TFP-2 features+KNN	63.23%	95.74%	88%
TFP-2 features+Decision Tree	31.18%	98.52%	83.97%
TFP-2 features+Linear SVM	70.81%	97.08%	90.97%
TFP-2 features+TreeBagger	76.25%	95.83%	91.27%
LBP features+LDA	72.96%	97.82%	92.03%
LBP features+QDA	78.40%	94.44%	90.70%
LBP features+KNN	78.97%	95.35%	91.53%
LBP features+Decision Tree	57.79%	90.65%	83%
LBP features+Linear SVM	90.41%	93.44%	92.73%
LBP features+ TreeBagger	89.98%	93.48%	92.67%
DWT+MFCC+KNN	62.00%	98.18%	92.60%
MFCC+Linear SVM	73.82%	97.16%	93.53%
DWT+MFCC+KernalSVM(poly)	73.39%	97.63%	93.87%
DWT+MFCC+Kernal SVM (rbf-100)	90.13%	92.03%	91.74%

V. CONCLUSIONS

In this paper, a new approach for NARW upcall detection has been proposed. In this approach, signals from passive acoustic sensors are directly fed into preprocessing, feature extraction and classification blocks without transforming them into spectrograms. This approach is not only more intuitive than those spectrogram-based methods given in the literature, but also produces better results in terms of both accuracy and speed. Different types of features have been investigated in this paper for call detection. It was shown that integration of MFCCs and DWTs produce the best performance among these tested with an experimental investigation.

It needs to be emphasized that in this research, the DWTs are applied to extract features of NARW upcalls, rather than to attenuate acoustic noises in the ocean. It has also to be pointed out that to the best knowledge of the authors, this is the first time DWTs together with MFCCs have been applied to extract features of NARW upcalls. The proposed method removes the need to transform these calls to the frequency domain, thus reducing the needed processing time by orders of magnitude. The method can also be used to detect and classify other mammal calls as the proposed procedure itself is generic.

ACKNOWLEDGMENT

The authors acknowledge the acquisition of the acoustic data made available by Cornell University.

REFERENCES

[1] D. K. Mellinger, K. M. Stafford, S. E. Moore, R. P. Dziak, and H. Matsumoto, "An overview of fixed passive acoustic observation methods for cetaceans," *Oceanography*, vol. 20, pp. 36-45, 2007.

[2] S. M. Van Parijs, C. W. Clark, R. S. Sousa-Lima, S. E. Parks, S. Rankin, D. Risch, and I. C. Van Opzeeland, "Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales," *Mar. Ecol. Prog. Ser.*, vol. 395, pp. 21-36, 2009.

[3] R.R. Reeves, B.D. Smith, E.A. Crespo, and G. Notarbartolo di Sciari, "Dolphins, Whales and Porpoises: 2002–2010 Conservation Action Plan for the World's Cetaceans," *IUCN/SSC Cetacean Specialist Group*, Chapter 4, 2003.

[4] C.W. Clark, "The Acoustic Repertoire of the Southern Right Whale, a Quantitative Analysis," *Anim. Behav*, vol. 30, pp. 1060-1071, 1982.

[5] D. Gillespie, "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," *Can. Acoust.*, vol. 32, pp. 39-47, 2004.

[6] D. K. Mellinger and C. W. Clark, "A method for filtering bioacoustic transients by spectrogram image convolution," *Proceedings of the IEEE: OCEANS '93*, vol. 3, pp. 122-127, 1993.

[7] I. R. Urazghildiiev, C. W. Clark, T. P. Krein, and S. E. Parks, "Detection and recognition of North Atlantic right whale contact calls in the presence of ambient noise," *IEEE J. Ocean. Eng.*, vol. 34, pp. 358-368, 2009.

[8] D. K. Mellinger, "A comparison of methods for detecting right whale calls," *Can. Acoust.*, vol. 32, pp. 55-65, 2004.

[9] M. Esfahanian, H. Zhuang, N. Erdol, and E. Gerstein, "Comparison of two methods for detection of North Atlantic Right Whale Upcalls," *EUSIPCO 2015*.

[10] M. Esfahanin, *Detection and Classification of Marine Mammal Sounds*, Ph.D. Dissertation, Florida Atlantic University, 2014.

[11] M.A. Roch, , M.S. Soldevilla, , J.C. Burtenshaw, , E.E. Henderson, & J.A. Hildebrand, 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*, vol. 121, p. 1737.

[12] P. J. Dugan, A. N. Rice, I. R. Urazghildiiev, and C. W. Clark, "North Atlantic right whale acoustic signal processing: Part I. Comparison of machine learning recognition algorithms," *Proc. IEEE: LISAT 2010*, 2010.

[13] S. Arivazhagan, W.S. Jebarani, and G. Kumaran, "Performance Comparison of Discrete Wavelet Transform and Dual Tree Discrete Wavelet Transform for Automatic Airborne Target Detection", *IEEE, International Conference on Computational Intelligence and Multimedia Applications,* pp.495-500, 2007.

[14] C.S. Burrus, R.A. Gopinath, and H. Guo, *Introduction to Wavelets and Wavelet Transform*, Prentice Hall, 1998.

[15] P. Motlíček, "Feature Extraction in Speech Coding and Recognition," *Research Report*, Oregon Graduate Institute of Science and Technology, pp. 1-50, 2002.

[16] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, 1995.

[17] B. Scholk, ¨ opf and A. J. Smola, *Learning with Kernels*, MIT Press, 2002.

[18] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.