Automatic Event Detection for Long-term Monitoring of Hydrophone Data

F. Sattar[a], P. F. Driessen[b], G. Tzanetakis[c], S. R. Ness[c], and W. H. Page[d],
[a] NEPTUNE Canada, University of Victoria, Canada
[b] Department of Electrical and Computer Eng., University of Victoria, Canada
[c] Department of Computer Science, University of Victoria, Canada
[d] Institute of Food Nutrition and Human Health, Massey University, New Zealand
email: farooks@uvic.ca, peter@ece.uvic.ca, grtzan@cs.uvic.ca, sness@sness.net,w.h.page@massey.ac.nz

Abstract

In this paper, we propose an efficient method for long-term monitoring of a wide variety of marine mammals and human related activities using hydrophone data. The proposed method uses a combination of a two-stage denoising process followed by a new event detection function that estimates temporal predictability. The detection function utilizes long-term and short-term predictions in order to detect various acoustic events from the background noise. The first stage of the denoising process uses temporal decomposition via Empirical Mode Decomposition to improve the correct detection rate, while the second stage uses Wavelet Packet spectral decomposition to reduce the false detection rate. Applied to event detection in NEPTUNE hydrophone recordings, the method demonstrates an accuracy of 95% and an F-measure of 94%.

1 Introduction

Long-term monitoring and proper interpretation of the variability of both the natural and anthropogenic components of the sound field in the ocean is essential for improving our understanding of the breaking waves, ocean atmosphere, as well as important questions associated with marine mammal conservation. Most existing monitoring systems use video information, but in some situations, such as monitoring events occurring under the ocean, sound information captured using hydrophones plays an important role. Acoustic (rather than visual) monitoring is used primarily for under water investigations because acoustic waves can travel long distances in the ocean. Visual monitoring is useful only for short range observations up to several tens of meters in depth at most, and is not suitable for monitoring whales or shipping which may be many kilometers away.

The term acoustic event means a short audio segment, which has rare occurrence, and is not predictable when and if it occurs and is of importance to the monitoring application. The task of detecting informative acoustic events from hydrophone data is difficult as this type of data is usually noisy and highly correlated. Much of the literature for unusual event detection has focused on video surveillance. For unusual events in audio, the paper [1] proposes a semi-supervised adapted Hidden Markov Model (HMM) framework, in which usual event models are first learned from a large amount of (commonly available) training data, while unusual event models are learned by Bayesian adaptation in an unsupervised manner; [2] robustly models the background for complex audio scenes with a Gaussian mixture method incorporating the proximity of distributions determined using entropy; [3] investigates a machine learning, descriptor based approach that does not require an explicit descriptors statistical model, based on Support Vector novelty detection. [4] applies optimized One-Class Support Vector Machines (1-SVMs) as a discriminative framework for sound classification.

Most existing acoustic event detection methods have high computational complexity and require large (preclassified by experts) training sets. In this paper we propose a method that uses a combination of a two-stage denoising process followed by a new event detection function that estimates temporal predictability. The detection function utilizes long-term and short-term predictions in order to detect various acoustic events from the background noise.

978-1-4577-0253-2/11/$26.00 ©2011 IEEE

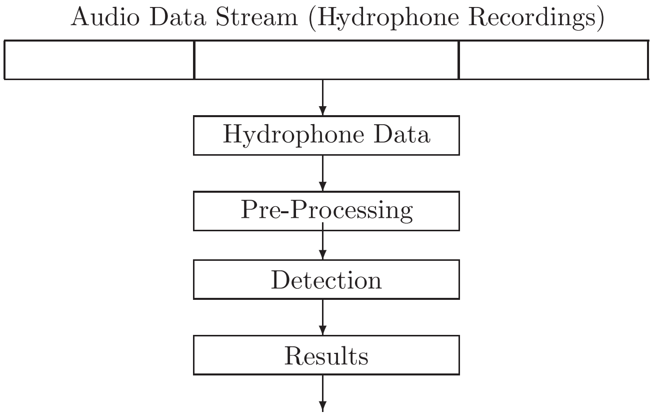

Figure 1. The overall block diagram of the proposed detection method.

2 Methodology

The overall block diagram of the proposed method is shown in Fig. 1. The input hydrophone data is preprocessed first by a two-stage denoising scheme introduced in section 2.2. Then the output signal is used as input for the proposed event detection scheme as presented in section 2.3.

2.1 Data

Different types of hydrophones are used for different tasks. For example, on the NEPTUNE Canada (http://www.neptunecanada.ca) observatory, an enhanced version of the Naxys Ethernet Hydrophone 02345 system is used, which includes the hydrophone element (rated to 3000m depth), 40dB pre-amplifier, 16-bit digitizer and Ethernet 100BaseT communication. This particular hydrophone is of high quality, and can be integrated into an existing underwater instrument package. The NEPTUNE hydrophone collects at a constant rate and can generate approximately 5.5 GB of data per day. The sampling frequency of the NEPTUNE data used is 96 kHz.

2.2 Preprocessing

The preprocessing is done by the following two stages as based on empirical mode decomposition (EMD) followed by wavelet packet transform.

2.2.1 EMD

The temporal decomposition is performed through empirical mode decomposition (EMD) by decomposing a signal into functions which form a complete and nearly orthogonal basis for the original signal. The functions, known as Intrinsic Mode Functions (IMFs), are sufficient to describe the signal, even though the are not necessarily orthogonal. In fact, the functions into which a signal is decomposed are all in the time-domain and of the same length as the original signal allowing for varying frequency content to be preserved over time. Obtaining IMFs from real world signals is important because natural processes often have multiple causes, and each of these causes may happen at specific time intervals.

In the EMD method, the original signal $x(t)$ can be represented in terms of IMFs as

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n \tag{1}$$

where $c_i(t)$ is the ith Intrinsic Mode Function and r_n is the residue and t represents the time samples.

The Empirical Mode Decomposition (EMD) method decomposes a signal into IMFs by an innovative sifting process [7]. A sifting process is performed to extract IMFs from the signal. The signal is processed iteratively in order to obtain a component which satisfies the conditions introduced in [7]. An intention behind these constraints on the decomposed components is to obtain a DC free single frequency component to guarantee a well-behaved Hilbert transform. It has been shown that the Hilbert transform behaves erratically if the original function is not symmetric with respect to the X-axis (i.e. no DC) or there is a sudden change in phase of the signal without crossing the X-axis [7]. The sifting process separates the IMFs with decreasing order of frequency i.e it separates high frequency component first and decomposes the residue obtained after separating each IMF till a residue of nearly zero frequency content is obtained. In a way, the sifting process in the EMD method may be viewed as an implicit wavelet analysis and the concept of the intrinsic mode function in the EMD method is parallel to the wavelet details in wavelet analysis.

2.2.2 Wavelet Packet Transform

The spectral decomposition is carried out by wavelet packet transform (WPT) or wavelet packet subspace (WPS) method [8, 9]. The reasons for choosing wavelet packets as well as wavelet decomposition/reconstruction are twofold: firstly, it helps to decorrelate the input signal from the correlated background noise by wavelet packet decomposition. Secondly, it can denoise the input signal by reconstructing the input signal from the selected subbands.

A wavelet packet is represented as a function,$\psi^i_{j,k}$, where i is the modulation parameter, j is the dilation parameter and k is the translation parameter.

$$\psi^i_{j,k} = 2^{-j/2}\psi^i(2^{-j}t - k) \quad (2)$$

Here $i = 1, 2, \cdots, j^n$ and n is the level of decomposition in wavelet packet tree. The wavelet ψ^i is obtained by the following relationships:

$$\begin{array}{l} \psi^{2i}(t) = \frac{1}{\sqrt{2}}\sum_{k=-\infty}^{\infty} h(k)\psi^i(\frac{t}{2} - k) \\ \psi^{2i+1}(t) = \frac{1}{\sqrt{2}}\sum_{k=-\infty}^{\infty} g(k)\psi^i(\frac{t}{2} - k) \end{array} \quad (3)$$

Here $\psi^i(t)$ is called as mother wavelet and the discrete filters $h(k)$ and $g(k)$ are quadrature mirror filters associated with the scaling function and the mother wavelet function [9], respectively. The wavelet packet coefficients, $c^i_{j,k}$ corresponding to the signal $x(t)$ can be obtained as

$$c^i_{j,k} = \int_{-\infty}^{\infty} x(t)\psi^i_{j,k}(t)dt \quad (4)$$

provided the wavelet coefficients satisfy the orthogonality condition.

The wavelet packet component of the signal at a particular node can be obtained as

$$x^i_j(t) = \sum_{k=-\infty}^{\infty} c^i_{j,k}\psi^i_{j,k}(t)dt \quad (5)$$

After performing a wavelet packet decomposition up to jth level, the original signal can be represented as a summation of all wavelet packet components at jth level as shown in Eq. (6).

$$x(t) = \sum_{i=1}^{2^j} x^i_j(t) \quad (6)$$

In fact, wavelet packet node entropy is used as a cost function to select the particular nodes/subbands for reconstruction. The entropy indicates the amount of information stored in the signal i.e. higher the entropy, more is the information stored in the signal and vice-versa. There are various definitions of entropy in the literature [10], such as energy entropy, Shannon entropy. The wavelet packet Shannon entropy at a particular node n in the wavelet packet tree of a signal is defined as $e_n = -\sum_k (c^i_{j,k})^2 log[(c^i_{j,k})^2]$ where $c^i_{j,k}$ are the wavelet packet coefficients at particular node of wavelet packet tree. Note that we have used here Shannon entropy although one can define his/her own entropy function if necessary.

The entropy has been calculated for every recordings for selecting the band to reconstruct it since it changes from one recording to another recording. So, by using entropy to automatically "select" the subband that contains the most information to detect the events, we do not have to use training data or to provide information about the frequency range of the events of interest.

2.3 Event Detection

The new event detection method proposed here is based on the idea of tracking the temporal predictability for time-varying signals. The temporal predictability function, $TP(n)$, can be defined as

$$TP(n) = log\frac{U[x(n)]}{V[x(n)]} \quad (7)$$

In Eq.(7), $V[x(n)]$ and $U[x(n)]$ refer to long-term prediction and short-term prediction, respectively, and n is the time samples.

The long-term and short-term predictions are performed by long-term autoregressive (AR) exponential averaging and short-term autoregressive (AR) exponential averaging defined by

$$V[x(n)] = \alpha_l V[x(n-1)] + (1 - \alpha_l)|x(n)| \quad (8)$$

$$U[x(n)] = \alpha_s U[x(n-1)] + (1 - \alpha_s)|x(n)| \quad (9)$$

where α_l and α_s are the exponential factors for the long-term prediction and the short-term prediction, respectively. In general, the exponential factor, α,

$$\alpha \approx \frac{1}{TFs} \quad \text{for} \;\; F_s >> f \quad (10)$$

where T is the time constant, Fs is the sampling frequency and f is frequency of the signal.

So, for a given Fs, we set the value of α_s based on the approximate time duration of the events, i.e. T_s. The value of the other parameter α_l is approximated based on the assumption that the noise is varying slowly, i.e. the corresponding time duration T_l would be larger than the time duration T_s.

Basically, $U[x(n)]$ is tracking the signal level of input $x(n)$ whereas $V[x(n)]$ is able to track the corresponding noise level. In order to make sure that the noise level follows the signal level and remain always below the signal level, we induce the following constraint:

$$V[x(n)] = \begin{cases} (1+\beta)V[x(n-1)] & \text{if } V[x(n-1)] \leq U[x(n)] \\ U[x(n)] & \text{if } V[x(n-1)] > U[x(n)] \end{cases} \quad (11)$$

where β is a small positive constant that controls how quickly the noise level adapts the changes for the noise characteristics.

The event signals are located based on the predictability function $TP(n)$. A binary detection function is then generated in the following way:

$$D(n) = \begin{cases} 1 & \text{if } TP(n) > Th \\ 0 & \text{if } TP(n) \leq Th \end{cases} \qquad (12)$$

where Th is used as threshold.
The output signal, $y(n)$, containing the detected events is obtained by masking the input signal $x(n)$ with the above binary detection function $D(n)$ as

$$y(n) = x(n) * D(n) \qquad (13)$$

where $'*'$ refers to point-wise multiplication.

3 Results and Discussion

3.1 Preprocessing

3.1.1 EMD

The results for temporal decomposition for Neptune data is illustrated in Fig. 2. An original input signal is illustrated in Fig. 2(a), whereas the corresponding decomposed signal using EMD for the decomposition level $i = 2$, is shown in Fig. 2(b). As we see in Fig. 2(a), the original input signal contains large spikes as outliers, which are significantly removed in the decomposed signal at level 2 as we have therefore used the second IMF signal. These spikes are probably caused due to echo locations as well as the effects of doppler acoustics.

3.1.2 Wavelet Packet Transform

The spectral decomposition has been done in the wavelet packet transform (WPT) domain. The wavelet packet transform enables us to reduce the noise as well as make it less correlated. An illustrative result is shown for the above processed Neptune data in Fig. 2(b). The Shannon entropy with the highest entropy value is used to choose this specific subband for reconstruction. Here, for example, the first subband out of 64 subbands (i.e. $=2^6$) at level 6, is automatically selected for reconstruction since it provides the highest entropy.

3.2 Event Detection (ED)

The event detection results for the Neptune hydrophone data (with a sampling frequency of 96 kHz) are illustrated in Fig. 3. The predictability function for the preprocessed signal in Fig. 2(c), is shown in Fig. 3(a). The corresponding binary decision function generated by choosing a threshold $Th = 0$ is depicted in Fig. 3(b). The output signal with the detected events based on the binary masking function is presented in Fig. 3(c) when the duration of the signal is 640 seconds. In Fig. 3(d), the zoomed version of the signal in Fig. 3(c) having a duration of 30 seconds is shown. An extracted event signal representing a whale call is illustrated in Fig. 3(e). The corresponding parameters used here are as T_s=1000 msec, T_l=2000 msec for which the parameters of α_s and α_l are calculated as 1.0417×10^{-5} and 5.2083×10^{-6}(with Fs=96000 Hz) and β=10^{-6}.

4 Performance Evaluation

For our performance evaluation we first incorporate the terms of precision and recall as defined by [11]

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (14)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (15)$$

where TP, FP, and FN represent the true positive, false positive, and false negative, respectively. Note that TP is defined as number of correctly detected events, whereas FP and FN are represented as number of wrongly detected events and number of missed detected events, respectively.

The following two quantitative measures are then obtained to evaluate the detection performance defined as

$$\text{F-measure}(\%) = 2(\frac{\text{precision}\times\text{recall}}{\text{precision+recall}}) \times 100 \qquad (16)$$

$$\text{Accuracy}(\%) = \frac{\text{Number of correctly detected events}}{\text{Total number of events}} \times 100. \qquad (17)$$

Table 1 presents the performance of the proposed scheme in terms of detection accuracy and F-measure on the labeled Neptune hydrophone dataset. It can be seen that the accuracy is quite high. The first row shows the performance when both the preprocessing by temporal decomposition and spectral decomposition are involved. While the detection performance for the preprocessed data only by temporal decomposition is presented in the second row. The third row shows the detection results without any preprocessing involved. As we can see the detection performance is improved due to the preprocessing. It can be noted that the correct detection rate(%) is increased by the temporal decomposition, while the false detection rate(%) has been improved by the following spectral decomposition.

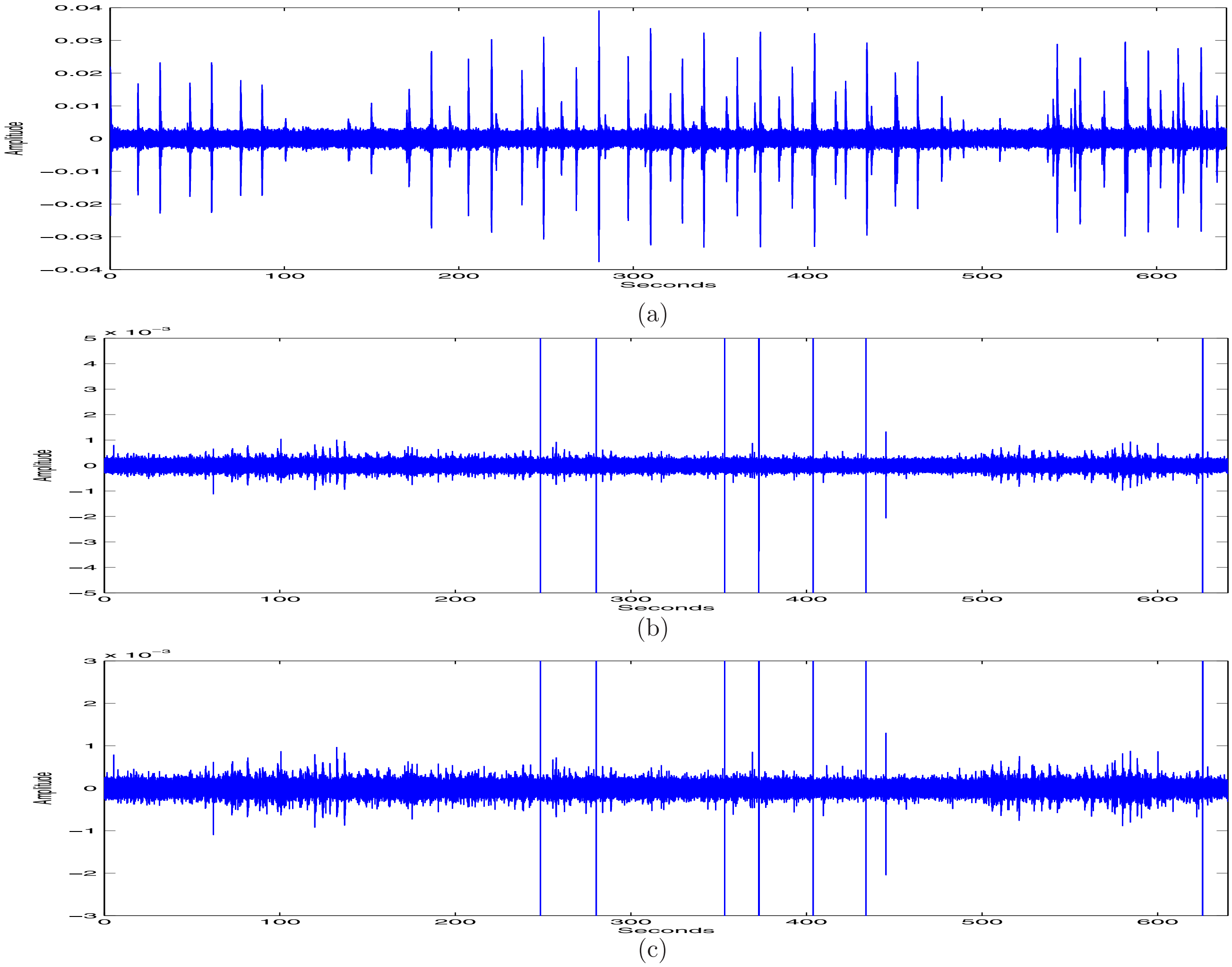

Figure 2. (a) The original Neptune data; (b) The results of temporal decomposition for the Neptune data at level 2; (c) The reconstructed signal after wavelet packet decomposition of the signal in Fig. 2(b).

Note that a total of 140 events are labeled for performance evaluation. The events which have been considered in this paper mainly consist of different kinds of whale calls with duration of 1-2 seconds long. There are total 12 recordings with each approximately 1 minute duration and 10-12 events/recording on average.

The proposed method has been compared with the common energy-operator based method [12]. It is found that unlike the presented method, the energy-operator based method is not able to identify the event signals from the high background noise. In Table 2, the detection performance of the energy operator based method is shown which is quite low in compared to the proposed scheme. A suitable threshold $Th = 10^{-4}$ is selected for the energy operator when the detection function is normalized to unit energy. Note that the energy operator based method totally fails when the preprocessing stages are not applied.

Table 1. The detection performance of the proposed scheme

Type of Processing	Accuracy	F-measure
EMD+WPT+ED	95%	94.3%
EMD+ED	89.29%	88.03%
ED	50.71%	57.49%

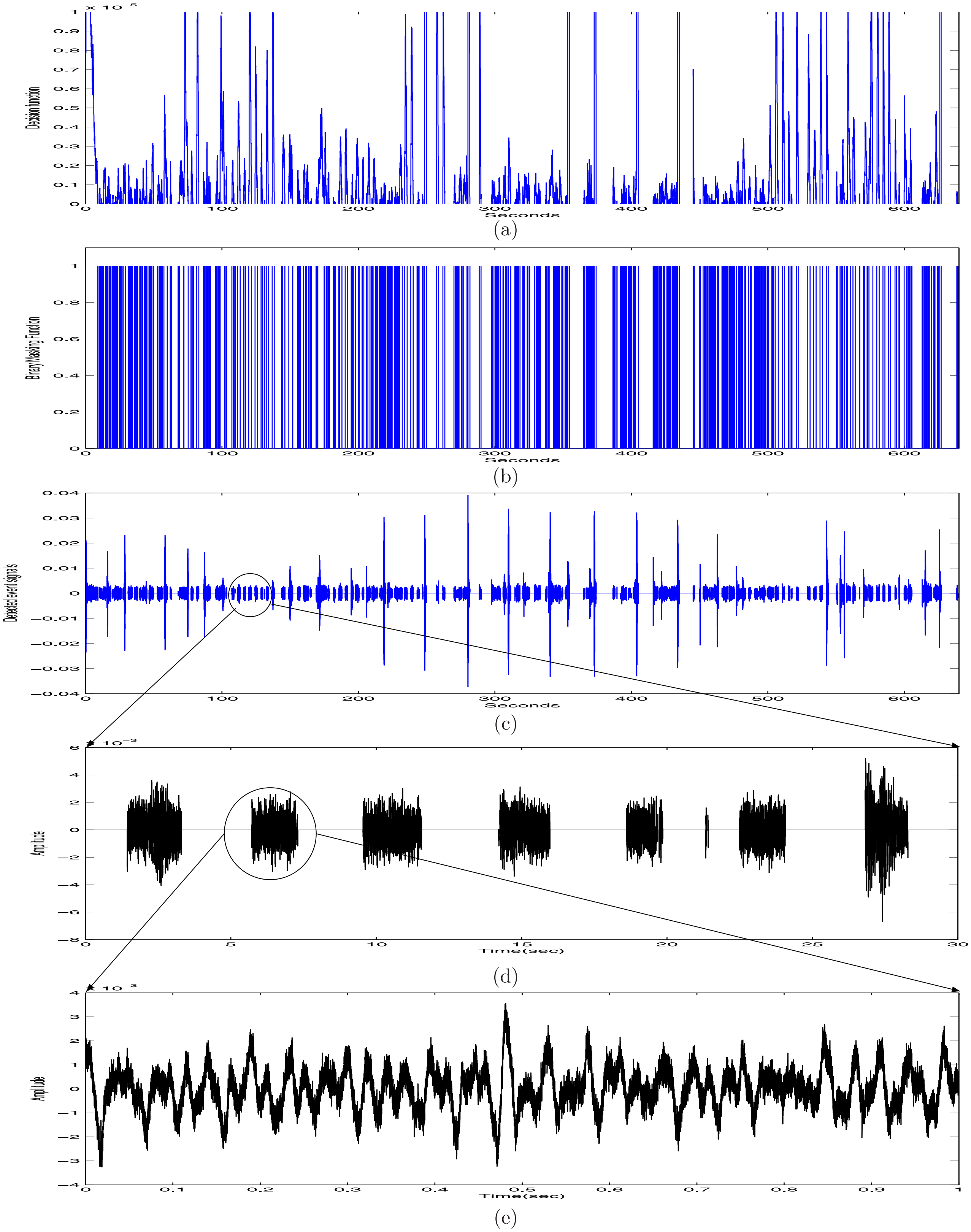

Figure 3. (a) The predictability function; (b) The binary decision function; (c) The output signal with the detected whale calls; (d) The zoomed version of the signal in Fig. 3(c);(e) One of the extracted event signal (a whale call).

Table 2. The detection performance of the energy-operator based method

Type of Processing	Accuracy	F-measure
EMD+WPT+ED	42.14%	34.10%

5 Conclusion and Future Work

An event detection method has been proposed for the long-term monitoring of marine mammals and human activities using hydrophone data. The presented method is based on a new detection function and is able to detect underwater acoustic events from the high background noise.

The presented scheme shows promising performance and outperforms the existing common energy operator based method. The ongoing work involves signal extraction for the detected events followed by useful feature extraction for classification of rare events in the highly noisy as well as correlated hydrophone recordings we are dealing with.

References

[1] V. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies", *Artificial Intelligence Review*, Vol. 22, No. 2, Oct. 2004, pp 85-126.

[2] D. Z. Daniel, G.-P. S. Bengio, and I. McCowan, "Semi-Supervised Adapted HMMs for Unusual Event Detection", in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition(CVPR'05)*

[3] M. Davy, F. Desobry, A. Gretton, and C. Doncarli, "An Online Support Vector Machine for Abnormal Events Detection", *Signal Processing*, Vol. 86, No. 8, Aug. 2006, pp. 2009-2025.

[4] A. Rabaoui, M. Davy, S. Rossignol, Z. Lachiri, and N. Ellouze,"Improved One-Class SVM Classifier for Sounds Classification", *IEEE Conf. on Advanced Video and Signal Based Sound Environment Monitoring (AVSS 2007)*, Sept. 2007.

[5] K. Yamanish, J. Takeuchi, and G. Williams, "Online Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms", *Proc. of KDD2000*, ACM Press, pp. 250254.

[6] V. Eva, P. Mat, V. Jozef, O. Stanislav, J. Jozef, and E. Anton "Detection and Classification of Audio Events in Noisy Environment" *Journal of Computer Science and Control Systems*, Vol. 3, No. 1, 2010.

[7] Huang et al, The Empirical Mode Decomposition Method and the Hilbert Spectrum for Non-linear and Non-stationary Time Series Analysis, *Proc. R. Soc. Lond* , 454, pp. 903-995, 1998. C.S.

[8] C.S. Burrus, R.H. Gopinath, and H. Guo, "Introduction to Wavelets and Wavelet Transforms, A Primer", *Prentice-Hall, Englewood Cliffs, NJ*, 1998.

[9] I. Daubechies, 10 Lectures on Wavelets, *Capital City Press*, 1992.

[10] R. Coifman, and M. Wickerhauser, Entropy based Algorithms for Best Basis Selection, *IEEE Trans. Information Theory*, 38, pp. 713-718, 1992.

[11] D. L. Olson and D. Delen, *Advanced Data Mining Techniques*, Springer Verlag, 2008.

[12] P. Maragos, J. F. Kaiserm and T. F. Quatieri, "On Amplitude and Frequency Demodulations using Energy Operators," *IEEE Trans. Signal Processing*, vol. 41, pp. 1532-1550, April 1993.

Acknowledgment
The authors would like to gratefully acknowledge the support from Neptune Canada under CANARIE project.