Automatic recognition of $T$ and teleseismic $P$ waves by statistical analysis of their spectra: an application to continuous records of moored hydrophones

Alexey Sukhovich*, Jean-Olivier Irisson*, Julie Perrot, and Guust Nolet

* Both authors contributed equally

Corresponding author: A. Sukhovich, Domaines Océaniques, UMR 6538, Université Européenne de Bretagne, Université de Brest, CNRS, IUEM, Place N. Copernic, F-29280 Plouzané, France. (sukhovich@univ-brest.fr)

Domaines Océaniques, UMR 6538,
Abstract. A network of moored hydrophones is an effective way of monitoring seismicity of oceanic ridges as it allows to detect and localize underwater events by recording generated $T$ waves. High cost of ship time necessitates long periods (normally a year) of autonomous functioning of the hydrophones, which results in very large sizes of the recuperated datasets. The preliminary but indispensable part of the data analysis consists in finding all $T$ wave signals. This process is extremely time-consuming as it is done by a human operator who visually examines the entire database. We propose a new method for the automatic signal discrimination based on the Gradient Boosted Decision Tree (GBDT) technique that uses the distribution

Université Européenne de Bretagne,
Université de Brest, CNRS, IUEM, Place N. Copernic, F-29280 Plouzané, France

\textsuperscript{2}Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, LOV, Observatoire océanologique, F-06230, Villefranche-sur-Mer, France

\textsuperscript{3}CNRS, UMR 7093, LOV, Observatoire océanologique, F-06234, Villefranche-sur-Mer, France

\textsuperscript{4}Géoazur, Sophia Antipolis, France.
of the signal power among different frequency bands as the discrimination
characteristic. We have applied the method to automatically identify the types
of the acoustic signals in the data collected by two moored hydrophones in
the North Atlantic. We show that the method allows to efficiently resolve
the signals of seismic origin, namely T waves and teleseismic P waves, with
a very small percentage of wrong identifications. In addition, a good iden-
tification rate for iceberg-generated signals is obtained. Our results indicate
that the method can be successfully applied to automatize the analysis of
other (not necessarily acoustic) databases provided that enough information
is available to describe statistical properties of the signals to be identified.
1. Introduction

The seismicity of the oceanic ridges and spreading centers is known to be characterized by a large number of earthquakes, most of which are not detected by land-based seismic stations due to their low magnitudes. To record these earthquakes, the common approaches are the use of ocean bottom seismometers (OBSs) [Kong et al., 1992; Wolfe et al., 1995; Bohnenstiehl et al., 2008]) or moored hydrophones [Dziak et al., 1995; Slack et al., 1999; Fox et al., 2001; Smith et al., 2002; Simão et al., 2010]. In the second case, several (four and sometimes more) hydrophones form a network which surrounds the area of interest. Compared to the OBSs, which are deployed over areas of limited extent, a hydrophone network can allow the surveillance of sections of the mid-oceanic ridges several hundreds of kilometers long. The hydrophones are positioned near the axis of the SOund Ranging And Fixing (SOFAR) channel which acts as an acoustic waveguide [Munk et al., 1995]. Seismic events along the mid-oceanic ridges, while of low intensity, frequently produce acoustic waves of appreciable magnitudes called T (short for Tertiary) waves [Tolstoy and Ewing, 1950]. These waves enter the SOFAR channel and, thanks to its waveguiding properties, can propagate long distances with almost no attenuation [Okal, 2008] until recorded by the hydrophones. When a T wave is detected by a network of hydrophones the coordinates of the spot where conversion of seismic to acoustic waves has taken place can be found [Fox et al., 2001]. During the last decade, hydrophone networks have detected many more events produced along the mid-oceanic ridges as compared to the land-based stations [Bohnenstiehl et al., 2003; Goslin et al., 2012]. The usefulness of hydrophone data is not limited to studies of the oceanic seismicity. For example, acoustic
signals generated by teleseismic $P$ waves were also observed in hydrophone records [Dziak et al., 2004]. The arrival times of these signals can be used in global seismic tomography whose progress is currently impeded by a lack of seismic data collected at sea [Montelli et al., 2004].

In all observation experiments with moored hydrophones, the monitoring of the acoustic pressure variation is done continuously. Because of high cost of ship time instrument installation and recuperation are separated by long time periods, normally a year or even longer in some cases, which results in very large datasets. The very first step in the analysis of the recuperated datasets consists in finding all recorded $T$ waves (or other signals of interest) which has to be done manually by a human observer visually inspecting one-by-one each day of recording. This preliminary part is extremely time-consuming due to large amount of the data. An automation of the signal identification is therefore highly desirable. Matsumoto et al. [2006] have proposed a simple algorithm to perform an automatic detection of $T$ waves by autonomous underwater freely-drifting robots. Recently, Sukhovich et al. [2011] have developed a probabilistic method for use on similar floats dedicated to record acoustic signals generated by teleseismic $P$ waves [Simons et al., 2009; Hello et al., 2011]. While showing promising results, these methods are not quite universal as they require separate tailoring of their parameters to the detection of the signals of a given type; once adjusted, they discriminate one signal type from the rest. In another approach, several authors used neural networks, which is a machine learning technique, for automatic picking of onset times of $P$ and $S$ waves [Dai and MacBeth, 1997; Zhao and Takano, 1999; Gentili and Michelini, 2006]. In this paper, we report on a new automatic method based on a different machine learning technique, the Gradient Boosted
Decision Trees [Breiman et al., 1984; Friedman, 2002], that is also adept at classifying multidimensional data. As compared to the neural networks, the GBDT technique is less opaque in the sense that one always has a complete information on each decision tree; when combined for all trees, this information allows to calculate for example the effect of a single discriminating variable or the joint effect of several variables. Similarly to the method of Sukhovich et al. [2011], our method uses the information on the distribution of the signal spectral power among different frequency bands; in contrast to Sukhovich et al. [2011], the use of the GBDT allows to classify several signal types simultaneously.

We have applied the new method to the data recorded by two moored hydrophones of the SIRENA network which operated in the North Atlantic during the years 2002 and 2003 [Goslin et al., 2012].

The paper is structured as follows. First, a detailed description of the method is provided. Afterward, a brief presentation of the hydroacoustic experiment SIRENA and the details on the data processing are given. The results are presented in the next section, which is followed by the conclusions.

2. Method description

We discriminate between signals of different types by using the differences in the statistical properties of their spectra. By “statistical properties” we mean how the spectral power of an ensemble of signals of a given type is distributed among a set of frequency bands. Thus, all signals are described by the same characteristics, which are derived from their spectral power distribution. These characteristics are used as variables in the GBDT procedure. The GBDT fits a statistical model which relates the characteristics of the sig-
nals to their type. Once the statistical model is defined, the identification of a signal of an
unknown type is performed by comparing its spectral power distribution with the model.

2.1. Signal processing

Spectral power distribution as a function of a frequency band is estimated with the help
of the discrete wavelet transform (DWT) similar to the procedure described by Sukhovich
et al. [2011]. For the sake of completeness, we provide here a brief description. The
wavelet transform is analogous to the Fourier transform (FT) in the sense that a given
signal is projected onto a space defined by a set of elementary functions called wavelets
[Jensen and la Cour-Harbo, 2001]. As a result, a set of wavelet coefficients is produced.
It is important to note that, contrary to sines and cosines which are elementary functions
of the FT, wavelets are functions of two variables, frequency and time. Thanks to this
fact, the DWT allows to obtain the information on the spectral content of the signal
both in the time and frequency domains, similarly to the short-time Fourier transform
(STFT, also known as spectrogram). The difference between the two is that the DWT is
non-redundant: it produces exactly the same number of wavelet coefficients as the length
of the signal to transform, while the STFT requires the calculation of FTs of overlapping
(usually by 50%) time windows. The DWT thus requires significantly fewer calculations.
Beside helping to speed up the analysis, this reduction of the amount of calculations is of
great importance when the analysis is done on the platforms with limited amount of power
available [Simons et al., 2009]. Another attractive feature of the DWT is the existence of
simple and easy-to-program algorithms for its calculation. In particular, we are using the
“lifting” algorithm [Sweldens, 1996].
Wavelet transforms use the notion of a “scale” rather than frequency. In some sense, the
wavelet transformation is also analogous to a filterbank analysis, with overlapping filters
covering different frequency bands [Jensen and la Cour-Harbo, 2001]. During wavelet
transformation, every iteration produces a set of wavelet coefficients for one scale, with
each scale corresponding to a particular frequency band. As a rule of thumb, each subse-
quent scale has a passband centered around a frequency that is half of the center frequency
of the previous scale. The first scale corresponds to the highest frequency band, which
takes an upper half of the total frequency band allowed by the sampling frequency $F_s$,
i.e. approximately between frequencies $F_s/4$ and $F_s/2$. The second scale corresponds to
the lower frequency band lying approximately between frequencies $F_s/8$ and $F_s/4$ and
so on for subsequent scales. Numerous wavelet bases exist and the most suitable one is
chosen for a particular problem at hand. We use a biorthogonal wavelet basis with two
and four vanishing moments for the primal and dual wavelets respectively, which is com-
monly abbreviated as CDF(2,4) [Cohen et al., 1992]. As was shown previously [McGuire
et al., 2008; Sukhovich et al., 2011], this wavelet construction provides a good compromise
between computational effort and the filtering performance of the wavelets.

A convenient way to visualize the result of the wavelet transformation is a scalogram,
which presents the absolute values of the wavelet coefficients in a time-scale plane. A
jump to Figure 5 shows representative signals of teleseismic $P$ and $T$ waves along with
their scalograms. The absolute value of a wavelet coefficient is proportional to the spectral
power of the signal at the corresponding moment of time and scale; we use this fact to
define an estimate $s_k$ of the signal’s spectral power at the scale $k$ as the average of absolute
values of the wavelet coefficients at this scale:
where $w^k_i$ is a wavelet coefficient, $N_k$ is the number of the wavelet coefficients and indices $i$ and $k$ number time and scale, respectively.

The detection of an arriving signal is ensured by monitoring the value of the ratio of short-term to long-term moving averages (STA/LTA algorithm) [Allen, 1978]. Equation (1) is evaluated for wavelet coefficients within a time window whose beginning is defined by an instance at which the value of the STA/LTA ratio rises over 2 (trigger threshold). After the trigger both moving averages continue to be updated and thus the STA/LTA ratio is affected by the triggering signal. The signal’s end is declared when the STA/LTA ratio drops below 1 (de-trigger threshold). These two values were found to work quite well in most cases: the lower trigger threshold would increase significantly the number of triggers due to ambient noise fluctuations while higher de-trigger threshold would not allow to capture the entire length of the signal. However, these values are not universal; some adjustments might be required if noise properties vary significantly from one site to another. The durations of time windows used for calculation of the short- and long-term moving averages were 10 and 100 seconds respectively.

From Equation (1), a set (or a vector) of scale averages $s_k$ is obtained. This set of values characterizes the absolute spectral power distribution among the scales and thus varies from one signal to another (of the same type) according to the signal-to-noise ratio. A more uniform characteristic is a relative power distribution $S_k$ found by normalizing the vector $s_k$ by its $L_1$ norm:
\[ S_k = \frac{s_k}{L_1}, \text{ where } L_1 = \sum_{k=1}^{K} s_k. \] (2)

In other words, normalization by \( L_1 \) norm gives the percentage of the total spectral power contained in each scale. The values of the \( S_k \) vector are signals’ characteristics introduced in the beginning of this section and used by the GBDT. Note that in the original implementation, Sukhovich et al. [2011] perform another normalization: each element \( S_k \) is divided by the corresponding element of the similar scale average vector calculated for the ambient noise record immediately preceding the STA/LTA trigger.

This second normalization by ambient noise means that a pure noise record with no signal should give elements \( S_k \) whose values are close to 1 for all scales. We have removed this normalization in order to improve the recognition rate of the iceberg-generated signals. As we shall see, these signals are very long and their amplitude fluctuates significantly with the time (Figure 6). As a result, it was common for the trigger to occur in the middle of the signal; thus, instead of the ambient noise, the record preceding the trigger contained the part of the same signal. The second normalization thus tended to “wash out” the distinctive features of iceberg-generated signals.

We illustrate the presented ideas on the actual case studied in this paper. Figure 1 shows the distributions of the scale averages \( S_k \) of four signal types observed in the records of two SIRENA hydrophones. From this figure it can be seen that statistical properties of \( T \) and teleseismic \( P \) waves are very different: while for \( T \) waves the power is mostly concentrated at scales 3 and 4, for teleseismic \( P \) waves most of the signal power is located at scales 5 and 6. For the remaining two types of signals, namely those generated by ships and icebergs, the most powerful scales are scales 3 and 4 respectively.
As will be explained in the section 2.2, the GBDT performs discrimination by comparing scale averages of the signals of different types at each scale. In this sense, Figure 2 represents the same information as Figure 1 but in the way the GBDT "sees" it. In each panel, the distributions of $S_k$ are compared for the signals of different types. The smaller the overlap between distributions is, the more discriminative and useful for classification the scale is. As expected from Figure 1, distributions of scale averages of $P$ waves are well separated from the rest of the signals at virtually all scales. The strong overlap of the distributions for $T$ waves, ship- and iceberg-generated signals at certain scales signifies that the discrimination of these signals will be more difficult than that of $P$ waves. It can also be seen that the scales 3 and 4 are likely to be the most discriminating ones as all four distributions overlap least at these scales.

2.2. Signal discrimination

To automatically identify the types of detected signals, we use the Gradient Boosted Decision Tree procedure. Before describing the GBDT, it is worth considering in some detail the basics of a classic decision (or classification) tree [Breiman et al., 1984]. A simplified single decision tree similar to those used in this paper is schematically described in Figure 3. A decision tree is essentially a set of rules according to which successive binary splits are performed; each binary split takes into account the value of one of the data characteristics. In our case, each signal is described by seven characteristics which are seven values of scale averages $S_k$. A classification of any dataset member (a signal in our case) is performed by passing it through the tree and distributing it along the branches according to the values of its characteristics. The tree itself is defined during the “training” phase, which is performed on a training dataset (i.e. a dataset whose
elements are already classified; in our case it is a dataset of the signals whose types has already been identified). A particular characteristic and its value at each binary split as well as the number of splits (i.e. the shape of the tree) are chosen automatically by a dedicated algorithm in such a way that the separation of the data results in groups of signals, each group containing as many signals of the same type as possible. Therefore, a decision tree is a supervised learning technique: the procedure first “learns” on a training set how to discriminate the data and then uses the statistical model (i.e. the tree) hence constructed to classify new unknown data. In the example presented in Figure 3, the training set consists of 10 signals of known types (4 T waves, 3 P waves, 2 ship- and 1 iceberg-generated signals). The constructed tree has two binary splits, which separate the signals according to the value of scale average at scales 3 and 5. Note that the succession of splits allows to take into account interaction between variables; in this example, the value of scale average at scale 5 is considered only when the value of scale average at scale 3 is \( \leq 2.1 \). Note also that the composition of any final group defines the likelihood for a classified signal to be of various types. As can be seen from Figure 3, there are 5 signals in the middle group: 4 T waves and 1 ship-generated signal with no P waves or ice-generated signals. Therefore an unknown signal, which ends up in this group, has a probability of \( 4/5 = 0.8 \) and \( 1/5 = 0.2 \) to be of a T wave and a ship-generated signal respectively; it has zero probability to be either a P wave or an iceberg-generated signal. As such, it is classified as a T wave.

If enough splits are performed, each of the 10 signals in the training set will be separated in its own “leaf”. Such a model would fit perfectly the training data but would lack generality (since minor peculiarities of the training set are unlikely to occur in other
datasets). A cross-validation procedure, whereby the training data is randomly split into subsets and the tree is grown with various combinations of the subsets, is used to determine which splits are robust, i.e. present for any subset of the training data. Such splits are likely to be also discriminant in other datasets and thus constitute the final statistical model.

The decision trees similar to that presented in Figure 3 are straightforward and easily readable, but they often have suboptimal predictive power. To improve the prediction, stochastic boosting procedures have proved successful [Friedman, 2001, 2002]. The general principle of boosting is to combine many successive small trees instead of using one large tree. The first tree is fit on the training data and divides it into few groups; it provides a first broad level of information (probabilities to be of each type in this case). The next tree is fit to the residuals of this first model and so on. The predictions of each successive tree are scaled by a constant, called the learning rate. It is usually much smaller than 1 and therefore decreases the influence of any individual tree and only allows patterns which occur in many trees to emerge from the ensemble, hence improving the generality and predictive power of the model. Stochasticity is introduced by randomly selecting a part of the data (50% in our case) to build each new tree. This also contributes to the generality of the model because patterns that are detectable even when only part of the data is used are main features of the dataset.

If too many successive trees are combined, unimportant peculiarities of the training set will be included in the model, which will hinder its generality and predictive power. In a regular decision tree procedure, cross-validation is used to decide when to stop growing the tree in order to keep the model general. In the GBDT procedure, the decisions should
be taken on when to stop growing each individual tree, how many trees to combine, and how much the model should learn from each tree. Maximum predictive power is achieved with: (i) individual trees large enough to account for interactions between characteristics but not too large to avoid extracting too much information in one tree, (ii) as many trees as possible, (iii) a very small learning rate for each tree. With 7 variables, 4 splits in each individual tree was a reasonable number and proved to be efficient. The learning rate was set to 0.001 which was small enough that a few thousand trees were needed to model the training dataset correctly. With this choice of parameters, the optimal number of trees was determined by 5-fold cross-validation. More precisely, the training data was randomly split into 5 equal pieces, a model was fit on 4 of them and several predictions were made on the 5th part, each prediction with an increasing number of trees. The discrepancy between the probabilities for each signal of being of each type predicted by the model and the actual type was quantified as the residual deviance. Initially, the residual deviance decreases sharply as more trees are added to the model, because each new tree helps to better classify the data. But at some point, additional trees only very marginally decrease the residual deviance and just waste computational resources. In our case, this point was reached between 3500 and 4500 trees.

All computations were carried out in the open-source software R (version 3.0.1, R Core Team [2013]) with packages gbm (version 2.1) for the GBDT [Ridgeway, 2013], reshape2 and plyr for data manipulation and automation [Wickham, 2007, 2011], and ggplot2 for graphics [Wickham, 2009].

3. Data
To test the method’s performance, we have used the data acquired during the hydroacoustic experiment SIRENA conducted in the North Atlantic from May 2002 to September 2003 [Goslin et al., 2004]. The hydrophone network comprised four instruments, which were installed on the western and eastern sides of an approximately 1000 km long section of the north Mid-Atlantic Ridge (Figure 4). The choice of this particular experiment for a verification of the method was dictated by the fact that this experiment’s data was already fully analyzed by the conventional approach of visual inspection. As a result, a catalog listing all the seismic events localized from the observed $T$ waves was available. The catalog allowed to minimize the amount of time spent on the preparation of the training dataset needed for the fitting of the statistical model (section 2.2).

The original sampling frequency of the SIRENA records was 250 Hz. To speed up the calculations, the records were filtered with an anti-aliasing low-pass filter with a corner frequency of about 40 Hz and then downsampled to 80 Hz. The low-pass filtering of the original data does not lead to the loss of information as the frequency content of the $T$ waves normally does not extend beyond 20 Hz. With this pretreatment, the first scale of the wavelet transform corresponds to the frequency band located approximately between 20 and 40 Hz (see section 2.1) while the last seventh scale would correspond to a frequency range between approximately 0.3 and 0.6 Hz. As can be seen from Figure 1, the seven-scale wavelet transform covers sufficient number of frequency ranges to evidence the differences in the spectra of the detected signals.

More specifically, we have applied the method to the records of the hydrophones S2 and S5 located at (42.72°N, 34.72°W) and (49.86°N, 24.57°W) respectively (Figure 4). The analysis covered the time period from June 1, 2002 to May 1, 2003. First, the
STA/LTA algorithm was run on the records (see section 2.1 for the STA/LTA parameters) which resulted in a total of 1570 and 1329 detected signals for hydrophones S2 and S5, respectively. In the next step, the types of the detected signals were manually identified in order to produce a training dataset for the GBDT. By calculating the expected arrival times of T waves generated by the events listed in the catalog and matching them with the times of the STA/LTA triggers, the vast majority of the detected signals were found to be T waves. Similarly, teleseismic P waves were identified by matching the times of the triggers with the arrival times of the seismic phases predicted for the model IASP91 [Kennett and Engdahl, 1991] for the events listed in the NEIC catalog during time the network was in operation. Signals whose trigger times were not matched by the predicted arrival times were examined visually. Some of these signals were found to be generated by passing ships. They normally lasted between 20 and 50 seconds and were easily identifiable from their spectrograms as can be seen from Figure 6. The last important class of the detected signals contains sounds generated by icebergs [Talandier et al., 2002; Dziak et al., 2010; Chapp et al., 2005; MacAyeal et al., 2008; Royer et al., 2009]. Quite generally, iceberg-generated signals are of very long duration and consist of a fundamental frequency accompanied by several (sometimes weaker) harmonics; an additional unmistakable property of such signals helping in their identification is the temporal variation of the emission frequencies (Figure 6). The rest of the detected signals were T waves whose origin events were not listed in the catalog (i.e. event localization was not possible since generated T waves were detected by less than three hydrophones) as well as several Pn waves. It was found that each of the Pn waves was followed by a high amplitude T wave arriving on average 200 seconds later. This small time separation suggests that each pair of Pn and T wave
signals was produced by the same nearby event. The spectral content of $Pn$ wave signals was very similar to that of $T$ waves. This similarity of spectral properties allowed us to consider $Pn$ and $T$ waves as signals of the same type from the point of view of the discrimination method. This also comes at no detriment to the future localization of the events since an identification of a $Pn$ wave would signify a simultaneous identification of a closely following $T$ wave signal. In the rest of the paper, the term “ensemble of the $T$ waves” includes also $Pn$ wave signals.

It should also be noted that both datasets contained numerous signals due to marine mammals vocalizations. However, being of very short duration, these signals were not detected by the STA/LTA algorithm whose parameters (above all the lengths of the short and long time moving time windows) were such that the STA/LTA ratio never exceeded the trigger threshold. Thus, the data analyses in geophysical and biological domains are conveniently decoupled by the very nature of the corresponding signals.

Table 1 provides a summary of the signals detected by each hydrophone.

4. Results

4.1. Evaluation of predictive power: self-prediction

A usual first step in an automatic classification analysis is the evaluation of the ability of a cross-validated model (which should therefore hold some generality) to re-predict the training dataset. Being built on the training set, the model is likely to be better at predicting this particular data than any other data. Therefore, the self-prediction results serve as a benchmark of the best possible performance the model can produce when applied to future independent datasets. Using the detected signals presented in Table 1 two separate models were fit to each dataset.
4.1.1. Hydrophone S2

The most discriminating characteristics of the model were scale averages at scales 3, 2 and 4 (these scale averages were present in the binary splits of 30, 30 and 23% of trees respectively). This could be expected from the distribution of scale averages shown in Figures 1 or 2: the signals of all types are quite well separated at these scales but overlap more at scale 7 for example. A convenient way to present the identification results is a confusion matrix (Table 2) which for all signals confronts their actual (columns) and predicted (lines) types in a single table. To know how many signals of a given type were predicted to be of another type, one needs to read the entry at the intersection of the corresponding column and line. For example, from Table 2 it can be seen that during the self-prediction test one T wave was predicted to be a P wave. When all the signals are identified correctly, only diagonal elements of the confusion matrix are non-zero. As all but one non-diagonal elements in Table 2 equal zero, the results of the self-prediction test for hydrophone S2 are almost perfect.

The performance of an identification method can also be described in terms of precision, recall and F1 score. These quantities summarize the more detailed information provided by the confusion matrix. Precision of the classifier quantifies how “pure” each predicted category is. It is computed as the number of true positives (i.e. correctly identified signals) divided by the total number of signals predicted in this category. Recall gives the fraction of signals of a given type that the model is able to correctly classify. It is defined as the ratio of the number of the true positives to the total number of signals of this type. One should note that a trade-off usually exists between precision and recall. Quite frequently, when a method correctly identifies all or most of the signals of a given
type (very high recall) it also ascribes a substantial number of signals of other types to
the same group thus lowering the precision. Conversely, when the precision is very high, it
is often because a method is too restrictive and thus many other signals of the same type
are rejected (low recall). The $F_1$ score combines both quantities into a single criterion;
the higher the value of $F_1$ score, the better is the performance of the method. The $F_1$
score is defined as follows:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$ (3)

Table 3 provides a summary of the confusion statistics for the self-prediction test of
hydrophone S2.

For each identified signal, the model actually predicts the probability to be of each
of the four types since each “leaf”, or final group, can contain signals of all four types
(Figure 3). The signal is identified to be of the type which has highest probability in
the group. To check how confident we can be in our classification, we compared the
four probabilities for each signal: the higher the probability corresponding to the true
type of the signal (as compared to the other three probabilities), the more robust is the
identification. Figure 7 confirms that the identification of signals of seismic origin is very
robust, with probabilities of being of the correct type mostly close to one and very little
possibility of confusion with other signals (since probabilities of being of another type are
near zero). The same holds for ship- and iceberg-generated signals. Only one $T$ wave
was identified as a $P$ wave; for this signal, the probabilities to be a $T$ wave and a $P$ wave
are low and very close, with that for a $P$ wave being slightly higher. This misidentified
signal decreases recall for $T$ waves and precision for $P$ waves. If the objective is to detect
as many $T$ waves as possible (i.e. high recall), a solution would be to concentrate only on the probability $p_T$ of being a $T$ wave (discarding the other three probabilities), set a threshold $t_T$, and consider any signal with $p_T > t_T$ as a $T$ wave. It should be remembered however that such an increase of the recall might come at the expense of the decrease of the precision as signals of other types but with a high enough value of $p_T$ would also be classified as $T$ waves. In case of hydrophone S2, the probabilities are different enough that even a relatively low threshold of 0.25 would ensure capturing all $T$ waves with very few false positives. If higher precision in the identification of $P$ waves is important, the solution would be to impose a more stringent criterion on the probability $p_P$ of being a $P$ wave. In addition to the requirement for $p_P$ to be the highest of the four probabilities, a threshold $t_P$ could be imposed such that any signal initially classified as a $P$ wave but with $p_P < t_P$ would be discarded from the $P$ wave group. Again, too high a threshold might also remove some other $P$ waves. As can be seen from Figure 7, a rather high threshold of 0.75 would eliminate the false positive due to a $T$ wave but also one true $P$ wave. These examples illustrate the trade-off between precision and recall; this trade-off has to be determined by the user depending on the relative importance of detecting absolutely all signals in a given group (high recall) vs. avoiding false positives (high precision).

4.1.2. Hydrophone S5

In case of hydrophone S5, the most discriminating characteristics were similar to hydrophone S2 but in a different order: binary splits based on the scale average values at scales 2, 3 and 4 are present in 37%, 36% and 12% of trees. This could also be expected from Figures 1 or 2: for example, signals are better separated at scale 2 for hydrophone S5 than for hydrophone S2. The confusion matrix for the self-prediction of the hydrophone
S5 data is presented in Table 4. The self-prediction results are by a tiny margin worse than those for hydrophone S2: three $T$ waves are misidentified as iceberg-generated signals and two ship-generated signals are predicted to be a $T$ wave and a sound emitted by an iceberg. The summary of the confusion statistics of the self-prediction test for hydrophone S5 is presented in Table 5.

4.2. Operational prediction: prediction from a subsample

The self-prediction tests provide a benchmark for the method’s performance. However, for the method to be useful, it should perform well with a statistical model created from a training set which is a small subsample of the entire dataset. The fewer signals are in the training set, the smaller is the amount of human effort required for their manual identification. Because we observed during the self-prediction test that the discriminating variables involved in the model are not exactly the same for hydrophones S2 and S5, we start by working on them separately before trying to use a single training set applicable to both hydrophones.

4.2.1. Selecting a representative data subset

As explained above, our goal is to select a small subset of the signals to fit the model and then use it to predict the types of the remaining signals. The important question is how to choose the signals for the training set. The training set needs to be representative of the entire dataset for the model to have acceptable predicting power. At the very least, it should contain signals of all types. However, because the vast majority of the signals detected by hydrophones are $T$ waves (Table 1), picking a portion of the signals at random is likely to result in a training set comprising only $T$ waves. Furthermore, even within $T$ waves, considerable variability in the distribution of scale averages exists.
(see the spread of the distribution of scale averages at scale 4 in Figure 2 for example); thus, the training set should also reflect this variability. To overcome the limitation of the naïve, random-choice approach, we have used an unsupervised classification technique to presort the signals in $k$ groups based purely on their characteristics (i.e. distribution of scale averages) without considering any *a priori* knowledge of their type. By construction, signals within each group look similar; signals from different groups look different from each other. As an example, the results of such unsupervised classification of the signals detected by hydrophone S5 are shown in Figure 8. The number of signals in each group is not constrained and therefore varies from one group to another. To build the training set, we randomly picked a *fixed* number $n$ of signals from each group. This resulted in a training set of $k \times n$ signals representing the *entire* variability of the complete dataset (picking the same *percentage* of signals in each group instead of a fixed number is in fact equivalent to randomly drawing the training set which thus will not be representative).

As the types of all signals in our datasets were known, we were able to check that the pre-classification indeed helps to separate $T$ waves from the much rarer signals ($P$ waves, ship- and iceberg-generated sounds) and that $T$ waves were themselves separated according to their varying properties (Figure 8). The outcome of the unsupervised classification was not very sensitive to the number of groups $k$. With as little as 3 groups, most $P$ waves are separated from the rest. With 4 groups, other rare signals are separated from $T$ waves. Increasing the number of groups provides further discrimination within $T$ waves and separates rare signals from each other. On the other hand, having too many groups might make them very small and extracting $n$ signals from each group would no longer be possible. A good trade-off between discrimination and sub-setting was found for $k = 10$. 

---

**Draft December 30, 2013, 7:39pm Draft**
The unsupervised classification algorithm we used was hierarchical clustering with Ward's aggregation criterion [Everitt et al., 2009]. The metric was the Euclidean distance between signals, computed from the “scaled” dataset: for each characteristic of the signal the mean is subtracted and the remainder is divided by the variance. Such scaling ensures that we give the same weight to all signal characteristics. Using Ward’s aggregation criterion allows to focus on the most marked differences between groups. Using other algorithms such as partition around medoids (k-medoids) or partition around means (k-means) gave very similar results.

Finally, one needs to determine what fraction of the entire dataset the training should use. A large fraction would result in a better prediction, because the training set would be more extensive. At the same time, a larger training set would also require more human effort on signal identification. A systematic test of fractions between 6% and 40% showed that a fraction of 10% was already sufficient to provide good identification rates of the seismic signals and that making the training set larger did not significantly improve the predictive power. Proportions smaller than 6% are not acceptable because the corresponding training sets did not comprise all observed signal types, even with the sub-sampling procedure described above, and were too small to fit enough trees.

4.2.2. Hydrophone S2

Since the signals for the training set are drawn randomly from the full dataset (divided beforehand into 10 groups using unsupervised classification as explained in the previous section), the identification results depend on the actual subsample drawn. To assess the effect of randomness, we have computed prediction statistics for 50 independently drawn subsamples. A summary of the results is presented in Table 6. The identification of $T$ and
P waves and iceberg-generated signals is quite good but ship-generated sounds are always misclassified. Even with such a small training set, P waves are almost always correctly identified (recall is virtually 100%). Almost all T waves are identified correctly (only four false negatives). Although signals of other types are also identified as T waves (about 1% of all signals identified as T waves are false positives), this result is quite acceptable, at least from the point of view of an operator who will use the predicted T waves to compose a catalog of seismic events. The composition of such a catalog favors rather higher values of recall (i.e. very few missed events) at the expense of precision. False positives will be ruled out during localization of the events and will not present a problem as long as they are not too numerous (which is obviously true in our case of only 18 false positives).

4.2.3. Hydrophone S5

Results obtained for hydrophone S5 (Table 7) are very similar to the results for hydrophone S2. Precision of the P wave identification is slightly worse but the recall is as good as for hydrophone S2. The confusion between T waves and other signals is also similar to that of hydrophone S2.

4.2.4. Hydrophones S2 and S5

When the data of hydrophones S2 and S5 are merged into a single dataset, a subsample of 5% gives a very similar performance (Table 8). Only one false negative occurs for P waves. As for T waves, nine false negatives were found, which nearly amounts to the sum of false negatives for T waves observed for each hydrophone separately. This means that, even though the self-prediction tests on each hydrophone suggested a different order of importance of the discriminatory variables, the fact that the same variables were discriminatory is sufficient to enable the construction of a common model for both
hydrophones. It should also be noted that in case of the merged set the human effort required for manual identification of the training set signals (5% of the total of 2899 signals or about 145 signals) represents half of that needed in case when identification is performed separately for each hydrophone (10% for each dataset or about 290 signals in total). In other words, the number of signals required to built an adequate training set for the merged dataset is approximately the same as that needed for each of the constituting datasets. This suggests that the quality of the identification depends on how well the training set represents the diversity of shapes of scale average distributions, rather than what fraction of the total number of available signals it contains. This is not surprising: once a distribution of the scale averages of a given shape is present in the training set, adding more signals with the same distribution shape does not bring new information. In our case, it seems that about 145 signals selected after the unsupervised classification were enough to cover most usual distribution shapes shown in Figure 8. A general advice would therefore be to pool as much data as possible into a single dataset before performing the classification. The larger total number of signals achieved by pooling data would also allow using a larger number of groups \( k \) in the unsupervised pre-classification, hence refining the different signal categories and possibly isolating and representing particularities of each hydrophone.

4.2.5. Source level of completeness

Another way to evaluate the method’s performance and usefulness is to estimate the source level of completeness \( (S_{Lc}) \) of the event catalog which would be obtained from the \( T \) waves identified by the method. The source level (SL) is a quantity which serves as an estimate of the magnitude of a hydroacoustic event, or in other words, the amount of...
energy released by the earthquake into the water column at the point of seismo-acoustic
conversion [Dziak, 2001]. It is measured in dB (with respect to micro-Pascal at 1 m)
and calculated from the magnitude of a $T$ wave by taking into account losses along the
propagation path and instrument response. The $SL_c$ is the minimum value of the SL at
which the logarithm of the cumulative number of events departs from a linear relationship.
The $SL_c$ is thus similar to the magnitude of completeness $M_c$ computed for a seismic
catalog by fitting the Gutenberg-Richter law. A lower $SL_c$ value corresponds to a more
complete catalog. The difference between the $SL_c$ obtained from the events identified by
our automatic method and the $SL_c$ derived from the complete catalog allows comparing
the method’s performance with that of a human analyst.

From the complete catalog, the value of $SL_c$ was estimated to be 209 dB; at this
value, the cumulative number of localized events differs from the one predicted from
Gutenberg-Richter law by 7.4%. Then, for each hydrophone, the cumulative numbers of
events as a function of the SL were calculated on the dataset predicted by the method,
for each of the 50 random realizations of the training set; these curves were found to
be virtually identical within each hydrophone. This is an important result as it further
underlines the small influence of the randomness in the choice of the training set on the
completeness of the seismic catalog. The performance of the method presented here is
therefore representative of what would be observed in a real-world scenario, when only
one training set would be drawn at random. This allowed to obtain an overall cumulative
number vs. the SL relationship for each hydrophone by averaging the curves found for
each subsample. Finally, the $SL_c$ values for each hydrophone were estimated from average
curves by finding the SL value at which the difference between the cumulative number and
the number predicted by fitting the Gutenberg-Richter law was equal to the percentage value found from the complete catalog, i.e. 7.4%. The $SL_c$ values thus estimated are 210.9 dB and 211.1 dB for hydrophones S2 and S5 respectively. These values are extremely close to one another and are only by about 2 dB different from the $SL_c$ estimated from the complete catalog. Thus, the catalog derived from the $T$ waves identified solely by our automatic method would be virtually identical to the one compiled manually.

5. Conclusions

A new method for the automatic signal identification is presented. The method utilizes differences in the statistical properties of the spectra of the signals of different types. Signal detection is performed with the classical STA/LTA algorithm. The information on the signal spectrum is obtained by calculating its wavelet transform while the identification of detected signals is realized with the Gradient Boosted Decision Tree technique. In this technique, a statistical model is first fit to a training set (i.e. a small subset an entire dataset) consisting of the signals whose types were manually pre-identified. The derived statistical model is then used to predict the types of the remaining unknown signals.

The method was applied to the 11 month long continuous records of two moored hydrophones which made part of a hydrophone network deployed during the hydroacoustic experiment SIRENA (North Atlantic). The STA/LTA algorithm was first run on datasets of both hydrophones and the types of all detected signals were manually identified. Both datasets contained signals of four different types: $T$ and teleseismic $P$ waves, ship-generated and iceberg-generated sounds. The results of the self-prediction tests with the GBDT (identification of the same data used to construct the statistical model) served as a benchmark of the best possible method performance. Using all detected signals...
as training sets, the self-prediction showed virtually perfect identification for both hydrophones. To simulate a more realistic case of an analysis of a new unknown dataset, a small portion of the signals was considered as the training set and then used for the identification of the remaining signals. The difficulty of choosing signals for the training set (which must be representative of the entire dataset) was overcome by using an unsupervised classification algorithm which allowed to split the ensemble of the signals into groups based solely on the statistical properties of their spectra. A collection of signals drawn at random from each group formed the desired representative training set. Choosing only 10% of the total number of the signals recorded by each hydrophone was sufficient to achieve high identification rates for the signals of seismic origin (telesismic $P$ and $T$ waves) which are the signals of the most interest in hydroacoustic experiments. In addition, the identification rate of iceberg-generated signals is comparable to that of seismic signals. The most difficult to identify were ship-generated signals. Furthermore, when the signals of both hydrophones were merged, the method provided similar performance with a training set making 5% of the combined dataset (approximately the same number of signals used to build training sets for each of the constituting datasets). In conclusion, our results demonstrate high potential of the presented automatic signal identification method, whose application is in principle not limited to the analysis of the hydrophone (i.e. acoustic) datasets. The method might thus prove useful for a community of geophysicists working with any large datasets and wishing to automatize the preliminary (i.e. signal identification) part of the data analysis.
Acknowledgments. Guust Nolet received financial support from the ERC (Advanced grant 226837). Alexey Sukhovich received financial support from the CNRS Chaire d'Excellence grant.
References


R Core Team (2013), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.


Royer, J.-Y., R. P. Dziak, M. Delatre, R. Chateau, C. Brachet, J. Haxel, H. Matsumoto, J. Goslin, V. Brandon, and D. R. Bohnenstiehl (2009), Results from a 14-month hydroacoustic monitoring of the three mid-oceanic ridges in the Indian Ocean, Geophysical


Table 1. Summary of the detections for hydrophones S2 and S5

<table>
<thead>
<tr>
<th>Signal type</th>
<th>Number of occurrences in S2</th>
<th>Number of occurrences in S5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$ waves</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>$T$ waves</td>
<td>1476</td>
<td>1224</td>
</tr>
<tr>
<td>Ship-generated</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>Iceberg-generated</td>
<td>63</td>
<td>71</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1570</td>
<td>1329</td>
</tr>
</tbody>
</table>

Table 2. Confusion matrix of the self-prediction test for the hydrophone S2 model

<table>
<thead>
<tr>
<th>Predicted type</th>
<th>Actual type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T$ waves</td>
</tr>
<tr>
<td>$T$ waves</td>
<td>1475</td>
</tr>
<tr>
<td>$P$ waves</td>
<td>1</td>
</tr>
<tr>
<td>Ship-generated</td>
<td>0</td>
</tr>
<tr>
<td>Iceberg-generated</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3. Confusion statistics of the self-prediction test for the hydrophone S2 model\(^a\)

<table>
<thead>
<tr>
<th>Type</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>precision</th>
<th>recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>T waves</td>
<td>1475</td>
<td>0</td>
<td>1</td>
<td>100%</td>
<td>99.9%</td>
<td>99.95%</td>
</tr>
<tr>
<td>P waves</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>94.4%</td>
<td>100%</td>
<td>97.12%</td>
</tr>
<tr>
<td>Ship-generated</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Iceberg-generated</td>
<td>63</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

\(^a\) The following abbreviations used: TP - true positives, i.e. correct identifications; FP - false positives, i.e. incorrect identifications; FN - false negatives, i.e. false rejections.

Table 4. Confusion matrix of the self-prediction test for the hydrophone S5 model

<table>
<thead>
<tr>
<th>Predicted type</th>
<th>Actual type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T waves</td>
</tr>
<tr>
<td>T waves</td>
<td>1221</td>
</tr>
<tr>
<td>P waves</td>
<td>0</td>
</tr>
<tr>
<td>Ship-generated</td>
<td>0</td>
</tr>
<tr>
<td>Iceberg-generated</td>
<td>3</td>
</tr>
</tbody>
</table>
Table 5. Confusion statistics of the self-prediction test for the hydrophone S5 model\textsuperscript{a}

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>precision</th>
<th>recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>T waves</td>
<td>1221</td>
<td>1</td>
<td>3</td>
<td>99.9%</td>
<td>99.8%</td>
<td>99.85%</td>
</tr>
<tr>
<td>P waves</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Ship-generated</td>
<td>10</td>
<td>0</td>
<td>2</td>
<td>100%</td>
<td>83.3%</td>
<td>90.89%</td>
</tr>
<tr>
<td>Iceberg-generated</td>
<td>71</td>
<td>4</td>
<td>0</td>
<td>94.7%</td>
<td>100%</td>
<td>97.28%</td>
</tr>
</tbody>
</table>

\textsuperscript{a} The following abbreviations used: TP - true positives, i.e. correct identifications; FP - false positives, i.e. incorrect identifications; FN - false negatives, i.e. false rejections.

Table 6. Confusion statistics in case of the identification of the signals detected by hydrophone S2\textsuperscript{a}

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>precision</th>
<th>recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>T waves</td>
<td>1471</td>
<td>19</td>
<td>5</td>
<td>98.7%±0.2</td>
<td>99.7%±0.1</td>
<td>99.2%±0.1</td>
</tr>
<tr>
<td>P waves</td>
<td>17</td>
<td>2</td>
<td>0</td>
<td>89.2%±4.2</td>
<td>99.9%±0.8</td>
<td>94.2%±2.4</td>
</tr>
<tr>
<td>Ship-generated</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Iceberg-generated</td>
<td>58</td>
<td>3</td>
<td>5</td>
<td>95.2%±2.8</td>
<td>92.1%±4.0</td>
<td>93.5%±1.8</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Confusion statistics is obtained from 100 subsamples randomly drawn from 10 groups formed by unsupervised classification procedure. For each signal type, we report the median numbers of true positives (TP), false positives (FP) and false negatives (FN) as well as the mean and standard deviation of the usual confusion statistics.
Table 7. Confusion statistics in case of the identification of the signals detected by hydrophone S5\textsuperscript{a}

<table>
<thead>
<tr>
<th>Type of Wave</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>precision</th>
<th>recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>T waves</td>
<td>1219</td>
<td>10</td>
<td>5</td>
<td>99.0%±0.2</td>
<td>99.6%±0.2</td>
<td>99.3%±0.1</td>
</tr>
<tr>
<td>P waves</td>
<td>22</td>
<td>6.5</td>
<td>0</td>
<td>77.0%±10.4</td>
<td>99.7%±1.4</td>
<td>86.5%±6.5</td>
</tr>
<tr>
<td>Ship-generated</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iceberg-generated</td>
<td>67</td>
<td>3</td>
<td>4</td>
<td>94.9%±1.8</td>
<td>92.8%±5.0</td>
<td>93.7%±2.4</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Confusion statistics is obtained from 100 subsamples randomly drawn from 10 groups formed by unsupervised classification procedure. For each signal type, we report the median numbers of true positives (TP), false positives (FP) and false negatives (FN) as well as the mean and standard deviation of the usual confusion statistics.
Table 8. Confusion statistics in case of the identification of the signals detected by hydrophones S2 and S5

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>precision</th>
<th>recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>T waves</td>
<td>2688</td>
<td>31</td>
<td>12</td>
<td>98.8%±0.3</td>
<td>99.5%±0.2</td>
<td>99.2%±0.1</td>
</tr>
<tr>
<td>P waves</td>
<td>38</td>
<td>6</td>
<td>1</td>
<td>85.5%±8.9</td>
<td>97.2%±2.6</td>
<td>90.6%±4.8</td>
</tr>
<tr>
<td>Ship-generated</td>
<td>0</td>
<td>0</td>
<td>26</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Iceberg-generated</td>
<td>126</td>
<td>9</td>
<td>8</td>
<td>92.8%±2.8</td>
<td>92.7%±5.3</td>
<td>92.6%±2.6</td>
</tr>
</tbody>
</table>

Confusion statistics is obtained from 100 subsamples, each subsample making 5% of the entire dataset (obtained by joining datasets for both hydrophones), randomly drawn from 10 groups formed by unsupervised classification procedure. For each signal type, we report the median numbers of true positives (TP), false positives (FP) and false negatives (FN) as well as the mean and standard deviation of the usual confusion statistics.
Figure 1. Dependence of the scale average $S_k$ on the scale number for the signals of different types (indicated on top of each panel) detected by hydrophones S2 and S5 (top and bottom panels respectively). For a given signal type, a panel shows scale averages of all detected signals; scale averages of the same signal are connected by the lines. The distribution of the scale averages within the same scale is represented by a violin plot. The extent of the violin plot gives the range of $S_k$ values while its width at a given $S_k$ value is proportional to the number of signals having similar value. The thickest part of the violin plot thus gives the position of the peak of the distribution.
**Figure 2.** Probability density functions estimated for all $S_k$ distributions (represented by violin plots in Figure 1). In contrast to Figure 1, which displays in each panel the distributions of scale averages of a single signal type for all scales, this figure compares the distributions of scale averages of all signal types at a single scale (indicated on top of each panel). Top and bottom panels correspond to the signals detected by hydrophones S2 and S5 respectively. To facilitate the visual comparison, the logarithm of $S_k$ is plotted along the horizontal axis and each probability density function is scaled to a maximum value of 1. The smaller the overlap between the distributions at a given scale is, the higher the discrimination power of this scale is.
Figure 3. The classification procedure of an unknown signal illustrated on a simple decision tree. The tree was constructed on the training dataset containing 10 signals of four different known types. An unknown signal is classified by passing it through the tree and distributing it along the branches according to the values of its characteristics (indicated in the right upper corner). The composition of the group into which the signal is distributed defines the likelihood for this signal to be of each of the four types. These probabilities are indicated in the lower right corner.
Figure 4. Bathymetric map of the Mid-Atlantic Ridge section surveyed during the SIRENA experiment. Stars indicate the hydrophones whose records were used to locate the events denoted by dots. Hydrophones S2 and S5 are in the lower left and upper right corners of the network respectively. White open triangles give the positions of the epicenters of large magnitude earthquakes listed in the NEIC catalog for the duration of the experiment. The inset gives a global view of the SIRENA network (with hydrophones again indicated by stars).
Figure 5. Representative acoustic signals of (left top panel) teleseismic $P$ wave and (right top panel) $T$ wave recorded by hydrophone S5 of the SIRENA network. The signals are downsampled to 80 Hz. Middle and bottom panels show signals’ scalogram and spectrogram respectively. Scalogram pixels give positions of wavelet coefficients in a scale-time plane. First scale corresponds to frequency range approximately between 20 and 40 Hz; second scale corresponds to the frequency range between 10 and 20 Hz, and so on. Both absolute values of the wavelet coefficients and the spectral density are measured in the units of the standard deviation (from the mean).
Figure 6. Representative acoustic signals produced by (left top panel) a passing ship and (right top panel) an iceberg recorded by hydrophone S5 of the SIRENA network. The signals are downsampled to 80 Hz. Middle and bottom panels show signals’ scalogram and spectrogram respectively. Both absolute values of the wavelet coefficients and the spectral density are measured in the units of the standard deviation (from the mean). The signals emitted by ships during the investigated period were quasi-monochromatic with a frequency of 8.5 Hz. As expected, the largest wavelet coefficients are located at the scale 3 (frequency range between 5 and 10 Hz approximately). On the contrary, iceberg-generated signals are extremely long and composed of several harmonics whose frequencies vary with time.
Figure 7. Probabilities to be of all four possible types (vertical axis) predicted by the method during the self-prediction test on the training dataset of hydrophone S2. Signals are split in four panels according to their actual type. To ease readability, in each panel signals are arranged in such a way that the probability of being of the correct type decreases monotonically (these probabilities are joined by a line). The highest probability among the four possible types is highlighted by a larger and darker symbol. The better the method’s performance, the higher is the probability corresponding to the actual signal type and the lower are the probabilities corresponding to other types. This should hold for as many signals as possible. It can be seen that the method performs very well for the signals of all four types, with the exception of one misidentification of a $T$ wave as a $P$ wave.
Figure 8. Unsupervised partition of signals detected by hydrophone S5 based on the scale averages through hierarchical clustering. This classification contributes to separating rare signals (P waves, ship- and iceberg-generated signals concentrated in groups 9 and 10) from more common T waves. It also separates T waves according to the intensity and distribution of scale averages: group 7 could be described as powerful signals with maximum energy in scales 3 and 4 while group 2 would be less powerful signals with maximum of the energy at scales 2 and 3. Picking 10 signals in each panel results in a collection of 100 signals more representative of the total variability of the entire dataset than picking 100 signals at random.