# Automatic fish sounds classification

Marielle Malfante, Jerome Mars, Mauro Dalla Mura, Cedric Gervaise

# Automatic Fish Sounds Classification

Marielle Malfante,[1, a)]
Jérôme I. Mars,[1] Mauro Dalla Mura,[1] and Cédric Gervaise[2]
[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP*, GIPSA-Lab, 38000 Grenoble, France*
**Institute of Engineering Univ. Grenoble Alpes*
[2] *Chorus, Fondation Grenoble INP, 38000 Grenoble, France*

(Dated: Accepted on Apr 13, 2018, DOI number: 10.1121/1.5036628)

The work presented in this paper addresses the issue of environmental monitoring. Specifically, it focuses on the use of acoustic systems for passive acoustic monitoring of ocean vitality for fish populations. To this end, various indicators can be used to monitor marine areas such as both the geographical and temporal evolution of fish populations. A discriminative model is built using supervised machine learning (random-forest and support-vector machines). Each acquisition is represented in a feature space, in which the patterns belonging to different semantic classes are as separable as possible. The set of features proposed for describing the acquisitions come from an extensive state of the art in various domains in which classification of acoustic signals is performed, including speech, music, and environmental sounds. Furthermore, this study proposes to extract features from three representations of the data (time, frequency, and cepstral domains). The proposed classification scheme is tested on real fish sounds recorded on several areas, and achieves 96.9% correct classification., compared to 72.5% when using reference state of the art features as descriptors. The classification scheme is also validated on continuous underwater recordings, thereby illustrating that it can be used to both detect and classify fish sounds in operational scenarios.

 [http://dx.doi.org(DOI number)]

[XYZ] Pages: 1–15

## I. INTRODUCTION

### A. Passive Acoustic Monitoring

The global ecosystem of the Earth is ruled and influenced by many factors, which include seas and oceans. Water covers around 70% of the Earth's surface, and it thus has a major role in the regulation of the global ecosystem. The economic value of the services provided by the seas is also substantial, at some $33 trillion per year according to (Costanza *et al.*, 1997).

Zones of shallow waters (depth $< 100$m) in coastal areas have therefore been widely studied since they have a key role in marine environments. In particular, sea-grass meadows that grow in shallow waters ($< 40$m) are of interest. At a local scale, sea meadows are the habitat of many fish species, as an environment that provides them with a source of food, and where they can develop their nurseries (Heck, 2003). At a larger scale, sea-grass meadows are considered to be the lungs of the oceans, in terms of their role in photosynthesis. These coastal areas are, by definition, closer to land, and their ecosystems are particularly exposed and vulnerable to anthropogenic stress. For these reasons, special attention needs to be paid to the evolution of such areas, which requires deployment of the necessary monitoring tools and devices. The development of an automatic monitoring system for sea-grass meadows is the main purpose of this study. Specifically, we focus on the monitoring of fish populations using passive acoustics.

The choice of considering acoustics signals as a proxy for monitoring of sea-grass meadows instead of other approaches, such as satellite imagery or field surveys, can be supported by three reasons. First, water is a particularly favorable environment for sound propagation. Conversely, satellite or airborne imagery is of limited use due to the difficulty in accessing the required underwater information. This arises from the limited visibility with passive optical systems, and the lack of penetration of electromagnetic waves through the water surface with active acquisition systems. Secondly, using an acoustic-based approach allows monitoring of large areas. Indeed, large amounts of data, and even real-time recordings, are relatively easy to acquire. As the cost of data collection campaigns is relatively affordable, their use can be considered for monitoring over wide areas. Finally, acoustic-based approaches are less invasive and costly than in situ surveys of the sea floor. Thirdly, the presence of biodiversity can be seen as an indicator of the vitality of an area since fishes and sea animals generally communicate through acoustics. In this context, analyzing fish sounds is a relevant choice for monitoring the biodiversity and vitality of a given coastal area. We propose here to use supervised *machine-learning* methods to build a classification model for automatic classification of fish sounds.

[a)] marielle.malfante@gipsa-lab.fr; Corresponding author.

To this end, the acoustic signals need to be represented as *feature vectors* before being assigned to one of several classes of interest.

Despite the need for automatic monitoring of seagrass floors, the existing tools based on acoustics are not adapted to handle the huge amounts of acquired data. Nevertheless, studies on automatic classification of more general acoustic data have been reported. Specifically, studies on environmental sounds (i.e., natural, animal, human sources), speech, and music have been conducted, and reviewed. Among the very few studies dealing with fish sounds classification, we can mention (Noda *et al.*, 2016; Vieira *et al.*, 2015) and (Sattar *et al.*, 2016) in which results are promising even if the tools are not all tested in an applicative context of underwater monitoring. The classification of bioacoustics data is being developed in the literature, such as the sounds of sea mammals (Zaugg *et al.*, 2010), bats, frogs (Chen *et al.*, 2012; Huang *et al.*, 2009), birds (Acevedo *et al.*, 2009; Fagerlund, 2007; Tyagi *et al.*, 2006) and other animals (Mitrovic *et al.*, 2006). However, depending on the field and data accessibility, the results have been relatively disparate. For instance, studies on speech recognition are relatively advanced, while automatic identification of musical instruments is at a more preliminary stage of development. However, to the best of our knowledge, no functional tool on automatic fish sounds classification has been reported to this date, and thus we propose the present study as an attempt to fill this gap.

The contributions of this study are as follows. (i) Development of a supervised classification architecture for automatic classification of fish sounds. (ii) Review of the literature regarding the representation of environmental sounds (i.e., natural, animal, human induced), speech, and music, to propose a large and comprehensive collection of features. (iii) Extraction of the considered features from three different representations of the acoustic signals: i.e., in the temporal, spectral, and cepstral domains, to yield a large set of descriptors that can capture different information, such as temporal evolution, predominant spectral frequency, and harmonicity of transient signals. (iv) Demonstrate the effectiveness of the proposed automatic classification system on a real dataset composed of acoustic acquisitions of fish sounds. We also conduct a comparative analysis on the features used, and provide some indications for the selection of the most relevant ones. (v) Testing of the proposed architecture on continuous recordings from various marine areas and at different times.

This paper is organized as follows. In Section I B, a state-of-the-art classification of acoustic signals is presented. In particular, we focus on the features used to represent the data. The fundamentals of machine learning are also summarized. Section II is devoted to the presentation of the architecture of the proposed automatic classification approach. In Section III, the data gathering campaign is presented, and the data used for this study are described. Then the experimental settings, along with the results of the various experiments, are presented in Section IV. Finally, a discussion is conducted in Section V, while the prospects and conclusions from this study are reported in Section VI.

### B. Related studies

In this Section, we review some studies from the literature on the classification of acoustic signals. In particular, we focus on the features used for classification of environmental sounds (e.g., natural, animal, anthropic), speech, and music, and on approaches that have been proposed for their classification. Before introducing the tools for the automatic analysis of acoustic data, we recall the fundamentals of machine learning and signal representations.

#### 1. Machine learning fundamentals

Machine learning can be described as a set of methods and techniques from artificial intelligence that are aimed at discriminating patterns (i.e., data samples[1]) into various semantic classes. The concepts of distance and similarity between patterns are crucial. The choice of the space used to represent the data is therefore also essential. Machine learning approaches aims at defining an automatic discriminative rule that can predict the most likely class of a new observation based for instance on a priori information on the classes as in the case of supervised approaches. Machine learning methods are currently widely spread, and are used in a broad range of applications, including medical imaging, image processing, speech processing, robotics, finance, and others.

Supervised learning algorithms employ a *labeled* dataset, in which each observation is associated with a thematic class (see sub-section III C for more details of the labeling process done in this work). In this scenario, supervised learning algorithms are trained on the labeled data leading to define a decision rule which minimizes the classification error on the available labeled examples. In many cases, the labeling step is carried out by hand, which is the main limitation of supervised learning approaches especially when data collections are huge. Nonetheless, supervised learning allows powerful prediction models to be built, and hence their extensive use throughout the scientific community. Numerous supervised-learning algorithms are used, including random forest (RF) (Breiman, 2001), support vector machine (SVM) (Boser *et al.*, 1992) and neural network (McCulloch and Pitts, 1943). For more details on machine learning algorithms, please refer to (Duda *et al.*, 2001) and (Friedman *et al.*, 2001).

In machine learning, as in many fields related to signal processing, the way signals are represented plays a fundamental role in the analysis. Features are extracted from the signals, shifting the representation of a numerical acoustic recording from a sequence of discrete samples in the time domain to a set of descriptors combined in a feature vector. The automatic classification rule is then established in the feature space, the domain in which patterns are represented where each dimension corresponds

to one feature. Reduction of the data dimensionality from the time signal length $n$ in the original space to the number of considered features $d$ in the feature space, in general can lead to better classification results. Theoretical results and reasons supporting this process can be found in (Duda *et al.*, 2001) or (Friedman *et al.*, 2001).

## 2. Classification of bioacoustic signals

By surveying the scientific literature, one can notice that a plethora of descriptors have been used as features for bioacoustic signal classification. These features come from different domains such as statistics, information theory, signal and image-processing.

For signal statistics, the spectral centroid and bandwidth have been used to classify frog sounds (Huang *et al.*, 2009). (Fagerlund, 2007) also used spectral centroid and bandwidth, and for recognition of bird species, they also included spectral roll-off frequency, spectral flux, and spectral flatness and duration. For the classification of whales and boats, (Zaugg *et al.*, 2010) used energy centroids, standard deviation, skewness, and kurtosis, all of which were computed from the time and frequency signals. (Acevedo *et al.*, 2009) used the minimal and maximal signal frequency ($f_{min}$ and $f_{max}$, respectively) in addition to the energy in different frequency bands to classify frog and bird songs. A threshold crossing rate was used by (Huang *et al.*, 2009) and (Fagerlund, 2007), and along the same lines, features based on area ratios above or below a given threshold were used by (Mitrovic *et al.*, 2006) for the classification of bird, cat, cow, and dog calls. Regarding underwater acoustics, (Noda *et al.*, 2016) introduces the use of Shannon entropy and call length for the recognition of fish sounds (in controlled environment though).

Low-level coefficients and descriptors issued from various transforms of the input signals have been also considered. For example, (Tyagi *et al.*, 2006) used dynamic time warping and spectral ensemble average voice for bird recognition. (Chen *et al.*, 2012) used multistage average spectrum for frog detection. Linear predictive coefficients have been also used in classification, as for calls from birds (McIlraith and Card, 1997) or humpback whales (Pace *et al.*, 2010). (Chesmore, 2001) used feature matrices from time-domain signal coding for insect and bird recognition.

Descriptors based on information theory have been used in (Han *et al.*, 2011), in which the authors implemented Shannon and Rényi entropies to classify frog calls, and in (Zaugg *et al.*, 2010) for the discrimination of whale and boat based on Shannon entropy.

Finally, some studies consider features extracted from spectrograms, which can be pragmatically considered as images allowing one to take advantage of the large set of image processing tools available. For bird vocalization retrieval, for example, (Dong *et al.*, 2013) extracted features from spectrogram images using ridge detection and points of interest. (Esfahanian *et al.*, 2014) also used image processing techniques for dolphin call classification. Bowhead whale are detected and localized in (Thode *et al.*, 2012), also using features extracted from the spectrogram.

Finally, features can also be learned from the signals instead of handcrafted, for example as done in (Sattar *et al.*, 2016) using principal component analysis for the classification of a given call of plainfin midshipman into three categories.

Pertaining to the classification procedure, several machine learning techniques have been employed such as SVM used in (Acevedo *et al.*, 2009; Fagerlund, 2007; Huang *et al.*, 2009; Mitrovic *et al.*, 2006; Noda *et al.*, 2016; Sattar *et al.*, 2016), neural networks (NN) are employed in (Chesmore, 2001; McIlraith and Card, 1997; Thode *et al.*, 2012) and (Zaugg *et al.*, 2010) while k-nearest neighbor (kNN) are considered in (Esfahanian *et al.*, 2014; Han *et al.*, 2011; Huang *et al.*, 2009; Noda *et al.*, 2016). Some studies also propose to use decision trees or linear discriminant analysis (Acevedo *et al.*, 2009), distance measurements (Dong *et al.*, 2013; Tyagi *et al.*, 2006) or k-means (Pace *et al.*, 2010). (Noda *et al.*, 2016) also uses RF.

## 3. Natural or human-induced sounds classification

For natural or human-induced sounds classification, very similar features to the ones previously presented have been used. Statistical descriptors were used by (Guo and Li, 2003) for multiple sounds retrieval, such as the frequency centroid, bandwidth in various frequency bands or pitch frequency and energy. To classify underwater mechanical transients, (Tucker and Brown, 2005) used a large number of perceptual features, which included energy standard deviation, skewness and kurtosis over the time or frequency axes. Image-processing techniques are also found for the description of these signals, such as by (Dennis *et al.*, 2011) for classification of various sound events. For low-level transforms, to classify acoustic noise radiated by boats, (Wang and Zeng, 2014) extracted features based on Bark wavelet analysis and Hilbert Huang transform. Linear predictive coefficients have also been used by (Couvreur *et al.*, 1998) for environmental noise recognition.

Regarding the learning algorithms, SVM are used in (Dennis *et al.*, 2011; Guo and Li, 2003) and (Wang and Zeng, 2014) along with distance measurements in (Guo and Li, 2003). kNN are also found in (Tucker and Brown, 2005) and Hidden Markov Models are tested in (Couvreur *et al.*, 1998).

## 4. Music classification

Automatic classification for music sounds is found in several various applicative domains (e.g., content retrieval, musical instrument identification, musical genre identification), although here again, different features have been used to describe signals of interest. Statistics features have been used, such as by (Eronen and Klapuri, 2000) and (Fujinaga and MacMillan, 2000) for musical instrument recognition and for timbre recognition. (Esmaili *et al.*, 2004) also used statistical features along with

entropy for classification of musical genre. Image processing methods have also been adopted, such as by (Yu and Slotine, 2009) for spectrogram texture extraction to identify various musical instruments, and (Deshpande *et al.*, 2001), for musical genre classification.

Learning algorithms used with musical data are similar to the ones used for the classification of biological, human or natural sounds. In particular, kNN is used in (Deshpande *et al.*, 2001; Fujinaga and MacMillan, 2000; Yu and Slotine, 2009), SVM is used in (Deshpande *et al.*, 2001) and linear discriminant analysis in (Esmaili *et al.*, 2004).

### 5. Speech classification and Mel frequency cepstral coefficients

Speech classification has been mainly carried out considering MFCC as features. MFCCs are based on a double Fourier transform or discrete cosine transform of the signal energy, thereby highlighting the harmonic properties of an acoustic signal. They are designed to describe sounds that are audible to humans, as they account for the perception of the human ear. MFCCs are very popular as features in the speech community, and they have been shown to be reliable for speech recognition and for speaker identification. Studies on various implementations of MFCCs have been conducted, such as by (Zheng *et al.*, 2001). The success of MFCCs in speech-related studies is such that they are now considered as a reference set of features for acoustic classification purposes in general. The state of the art in automatic classification of fish sounds is very limited but both (Vieira *et al.*, 2015) and (Noda *et al.*, 2016) have used MFCCs for fish call recognition or fish individuals classification. In bioacoustics, MFCCs have been used by (Bedoya *et al.*, 2014) for anuran sounds classification, by (Fagerlund, 2007) for bird call recognition, and by (Tyagi *et al.*, 2006) for bird species recognition. (Pace *et al.*, 2010) used MFCCs for humpback whale identification. (Lee *et al.*, 2006) and (Clemins and Johnson, 2006) modified MFCCs to fit their data, and they used them to classify frogs and crickets respectively, and for land mammal call identification. In acoustics, MFCCs were used by (Guo and Li, 2003) and (Dennis *et al.*, 2011) for multiple sounds identification, and by (Lim *et al.*, 2007) for identification of underwater acoustic transients. They were also used by (Wimmer *et al.*, 2010) for the identification of environmental sounds, and by (Márquez-Molina *et al.*, 2014) for aircraft take-off noise classification. MFCCs were also used to describe signals to distinguish speech from music from nonvocal sounds by (Foote, 1997), for musical instrument recognition by (Eronen and Klapuri, 2000), and for musical genre classification by (Deshpande *et al.*, 2001).

Similarly to other applications, learning algorithms involved in those studies are mainly based on SVM (Dennis *et al.*, 2011; Deshpande *et al.*, 2001; Fagerlund, 2007; Guo and Li, 2003; Noda *et al.*, 2016), neural networks (Márquez-Molina *et al.*, 2014), distance measurements (Guo and Li, 2003; Lim *et al.*, 2007; Tyagi *et al.*, 2006), k-means (Pace *et al.*, 2010), kNN (Deshpande *et al.*, 2001; Noda *et al.*, 2016), linear discriminant analysis (Lee *et al.*, 2006), hidden markov models (Clemins and Johnson, 2006; Vieira *et al.*, 2015) or RF (Noda *et al.*, 2016). Fuzzy classifiers were also used in (Bedoya *et al.*, 2014) and tree bases quantizer in (Foote, 1997).

## II. METHODS AND TOOLS

We now present the proposed architecture for automatic fish sound classification. The architecture we propose relies upon a labeled dataset of observations (i.e., it is a supervised approach) and includes four different steps: (i) preprocessing of the signals; (ii) extraction of features from the acquisitions; (iii) learning; and (iv) testing of the model. An extra step of feature selection can also be included. We detail each step in the following.

### A. Preprocessing

Preprocessing is used to condition the signals before the feature extraction phase. First, we consider as an observation a signal of length $\Delta_t$, which is filtered in frequency. The frequency bandwidths to be used are the same as the one planned in the labeling stage that will be fully detailed in Section III C: 50 Hz to 450 Hz, and 500 Hz to 900 Hz. One temporal window of length $\Delta_t$ therefore leads to two observations. The observations can be down-sampled without consequences to reduce the computational burden since only low frequencies are of interest here ($< 1kHz$). In addition, signals undergo a normalization so that each observation has unit energy in the temporal interval of duration $\Delta_t$. By doing this, classification becomes less dependent on the distance of fishes from the recording device. In other words, after energy normalization the signals are classified depending on their shapes, and not on their energetic content which would be a dominating feature.

### B. Features extraction

As explained in Section I B 1, the choice of features used to represent the data can have a major impact on the classification results. In this study, we used 84 features that were presented or inspired by works appeared in the literature and previously detailed in Section I B. The novelty of the proposed representation lies in: (i) The large set of features considered simultaneously, which is large, and is therefore more likely to capture most of the signal properties needed to discriminate the signals into their classes. (ii) The features are also general shape descriptors, which means that they can be used for other purposes, including classification of different transient signals. We purposefully gathered relatively general features instead of designing features that would be specific to an application or a particular family of sounds. By doing this, the features set can be used on other datasets (Malfante *et al.*, 2017). (iii) Finally, we extract the features from several signal representations.

Specifically, the feature vector of an observation is a concatenation of features extracted from:

- The time domain $x(t)$, to describe the waveform (`Time` feature set);
- The frequency domain $X(f) = \mathcal{TF}\{x(t)\}$, to describe the spectral content (`Frequency` feature set). In practice, Fourier transform is computed on $n$ points, with $n = 10400 \cdot 0.5$ as the observation length;
- The spectrum of the frequency domain, also referred as the *Cepstral* domain in the speech community $\mathcal{X}(q) = \mathcal{TF}\{X(f)\}$. This domain shows the periodic properties of a spectrum, which represents the harmonicity of the signal (`Cepstral` feature set).

The full list of the features is given in Table I. These features describe the observations and their properties. For example, we use entropy measurements, shape descriptors, and statistical moments, such as standard deviation, skewness or kurtosis, that define the spread, asymmetry, and flatness of a signal. Skewness computed in time describes the asymmetry of the time signal compared to the Gaussian distribution. Computed in frequency, it measures asymmetry of the signal spectrum and in the Cepstral domain, it describes the asymmetry of the representation underlying the harmonic properties of the data.

### C. Model training

Once extracted, the labeled feature vectors are used as input for the learning algorithm. This stage is referred to as learning or training, and, by learning a decision rule from the input data, it produces a classification model. In this study, we used SVM and RF algorithms as classifiers. We recall that, SVM aims at finding an hyperplan which optimally separates the labeled data into their classes. A kernel function can also be used to transform the input feature space to a space of larger dimension where the data can be linearly separable (Boser *et al.*, 1992). The RF algorithm is based on an ensemble of binary decision trees as weak learners. Classification is then obtained by majority voting on the predictions guiven by the ensemble of classifiers (Breiman, 2001). If the feature space has been efficiently defined for the data, the classification results should not be significantly influenced by the learning algorithm choice. Both algorithms have hyperparameters that need to be tuned: choice of the kernel and its parameters, and cost parameter $C_{SVM}$ for SVM, and number of trees and number of variables for RF.

### D. Model testing

In this study, we use two different procedures to test our models. The first one is known as cross-validation: from the N signals of the dataset, $\alpha N$ ones with the learning rate $0 < \alpha < 1$ are used to train the model, and the remaining $(1 - \alpha)N$ are used to test it. This process is repeated 50 times with random realizations of the training and testing sets, to ensure statistically robust results. This procedure is used to estimate the model performances, but also to set the hyperparameters to their optimum values for this application. Class by class accuracy and overall accuracy are used to measure the model performances.

$$Accuracy_i = \frac{\#_{TruePredictionsForClass\ i}}{\#_{ObservationsOfClass\ i}} \tag{1}$$

$$Overall\ accuracy = \frac{\#_{TruePredictions}}{\#_{Observations}} \tag{2}$$

To ensure that the learned model does not overfitting the training data, we also check performance on continuous underwater recordings. To do so, the recordings are continuously processed in the two frequency bandwidths o(i.e., 50-450 Hz and 500-900 Hz), with a sliding window of duration $\Delta_t$. To reduce the computational burden no overlap is considered between consecutive windows. For each observation, the model outputs the probabilities of the observation belonging to each of the classes. Those probabilities are then thresholded: if the probability associated to the most likely class is above the threshold, the predicted class will be retained as a reliable decision. If below the threshold, the observation will be considered as `Unknown` since it is significantly different from the known labeled samples. The validation process is carried out by domain experts by reviewing *a-posteriori* the classification results. To evaluate the model in this second configuration, we use the class by class and overall accuracy, but also the precision index which evaluates the false alarm rate.

$$Precision_i = \frac{\#_{TruePredictionsForClass\ i}}{\#_{PredictedAsClass\ i}} \tag{3}$$

Confusion matrices, from which all metrics are deduced are also presented to study the repartition of the errors and limitations of a model. A confusion matrix is a square matrix of size equal to the number of classes considered, in which each column shows for a given class $c$ with $1 \leq c \leq C$, as the number of correct and incorrect predictions.

### E. Features selection

The proposed architecture can also be extended to address the feature selection issue. In this study, we use a forward selection method: features are ranked by importance based on their weight in the RF model (see (Breiman, 2001) for more details of this process). Once ranked, the most important features can then be selected to form a subset that will be used to represent the data. This stage can be particularly significant when computational power is limited; e.g., in real-time applications. As previously explained, the feature set proposed in this study can be used for other applications; however, the subset of selected features is highly dependent on the

dataset, and is therefore meaningful for the considered application only. Using feature selection tools gives two interesting perspectives to our study: first, it helps in the estimation of the number of features needed to reach a given accuracy threshold; and secondly, it can be used as a tool to analyze the data. In particular, the physical meaning associated to the most important features can often lead to certain knowledge about the data and what discriminates them into their classes.

## III. DATASETS

### A. Data acquisition campaign

The data used in this study for the experimental analysis were collected in August 2014 in France during the SEACOUSTIC2014 campaign (Lossent *et al.*, 2015). The project aimed to collect data to address three issues: (i) How to determine the vitality of underwater areas; (ii) How to evaluate the anthropogenic stress on underwater areas; and (iii) To study the link between vitality of a given underwater area and the anthropogenic stress it faces. The campaign was based at the STARSEO station, near La Pointe de la Revelatta in Corsica, France (Mediterranean Sea). For more details on the SEACOUSTIC2014 project, please refer to (Lossent *et al.*, 2015).

The study presented here is related to issue (i); namely, the development of tools to determine the vitality of underwater areas. The data used to test and evaluate our system are area specific, which means that self-sufficient recording devices were fastened to the sea floor and left to record continuous signals. Specifically, between 1 day and 3.5 days of continuous recordings were collected from various marine areas. Each marine area was characterized by its depth and its sea floor (i.e., meadow, rock, sand, as given in Table II) and the development of automatic models to classify fish sounds will help in the analysis and characterizing of these areas. Hydrophones (HTI92 WB) and recorders (SDA14; RTSYS) were used, which provided the signals coded on 24 bits at 256 kHz (note, anti-aliasing filtering was applied). We here stress that the recordings used in the present study were not labeled, and no ground truth was available regarding their content. The labeling task was carried out specifically for the present study, and was conducted manually by an expert who reviewed and labeled the content of part of the recordings. More details are given in Section III C.

### B. Fish sound signals overview

The underwater recordings collected sounds from three different sources: animal (biophony), environmental (geophony), and human (anthrophony). Together, these formed the soundscape of a given area. In this study, we focus on biophony, and more specifically, on fish sounds.

Although a direct relationship between fish sounds and individual fish or fish species has not been established to date, it is known that specific fish sounds can be associated with specific behaviors, if the fish species is known, as determined in the fish biology literature (Amorim *et al.*, 2004; Amorim, 2006; Dos Santos *et al.*, 2000; Mann *et al.*, 2008; Parmentier *et al.*, 2006; Thorson and Fine, 2002). An established terminology that refers to the different fish sounds is still lacking, as well as any universally accepted correspondence between fish sounds and fish behaviors) (Amorim, 2006). For this reason, we chose to name (arbitrarily) the fish sounds in the acquisitions based on their qualitative characteristics. For the various recordings, four different types of fish sounds (*classes*) that showed distinctive acoustic signatures were recognized by the experts. Spectrograms of representative signals along with descriptions of the four classes are shown in Figure 1, and are hereafter detailed as:

`Impulsions`: Emissions of short duration that are separated by lags of the order of seconds.

`Drums`: Periodic pulse trains (around at least 15 pulses) that last for at least 20 s in most of the recordings.

`Roars`: Wideband signals in their frequency content, which are usually very energetic. These also last between about 10 s and 30 s, and occur at a frequency of around two a minute.

`Quacks`: Short signals with a harmonic structure that are often present (up to five occurrences per second). `Quack` sounds are recognizable by their similarity to frog or duck sounds.

These four classes were identified by the experts as clearly defined and matching the literature. Fish sounds that did not fit this nomenclature were particularly rare in these recordings. Alternatively, it would also be possible to define sub-classes: `Drums`, for example, could be sorted depending on the pulse frequency, and `Impulsions` depending on their frequency, which can vary significantly. Such sub-classes would more likely to be related to species or identification of individuals (Amorim, 2006; Mann *et al.*, 2008).

Sounds different from those belonging to the four above mentioned classes can also be spotted in the recordings. They were mainly due to ambient and anthropic noise. Such sounds are referred to as `Unknown` when a definite structure cannot be identified by the experts, or as `Background` when background noise dominates. The four fish sound classes are referred as the *positive* classes, while `Background` and `Unknown` represent the *negative* classes. More details on how those two classes were handled will be given with the results in Section IV.

According to the literature, the sounds named here as `Impulsions` and `Drums` appear to be related to antagonistic behaviors of fish (Amorim *et al.*, 2004; Dos Santos *et al.*, 2000; Parmentier *et al.*, 2006; Thorson and Fine, 2002), while `Roars` and `Quacks` would be produced during courtship (Dos Santos *et al.*, 2000; Thorson and Fine,

TABLE I. Feature set for a generic numerical signal $s[i]_{i=0}^{n}$ composed of $n$ discrete samples and which can be defined in different spaces such as the time $x(t)$, frequency $X(f)$ or cepstral $\mathcal{X}(q)$. $E$ represents the signal energy. Features references are referred 'T', 'F' or 'C' depending on their computation domain respectively being time, frequency of cepstral. Features in bold font are the most valuable features (see Fig 3).

| Feature | Definition | Used in | Ref. |
|---|---|---|---|
| Centroid | $\frac{1}{E}\sum_i i \cdot E_i$ | (Fagerlund, 2007; Huang *et al.*, 2009) | T1, **F1**, C1 |
| RMS bandwidth | $RMS_i = \sqrt{\frac{1}{E}\cdot\sum_i i^2 \cdot E_i - \bar{i}^2}$ | (Tucker and Brown, 2005) | T2, F2, C2 |
| Standard deviation | $\sigma_s = \sqrt{\frac{1}{n-1}\sum_i (s[i]-\mu_s)^2}$ | (Tucker and Brown, 2005) | T3, F3, C3 |
| Skewness | $\frac{1}{n}\cdot\sum_i \left(\frac{s[i]-\mu_s}{\sigma_s}\right)^3$ | (Zaugg *et al.*, 2010) | T4, **F4**, **C4** |
| Kurtosis | $\frac{1}{n}\cdot\sum_i \left(\frac{s[i]-\mu_s}{\sigma_s}\right)^4$ | (Zaugg *et al.*, 2010) | T5, F5, C5 |
| Mean skewness | $\sqrt{\frac{\sum_i (i-\bar{i})^3 . E_i}{E.RMS_i^3}}$ | (Tucker and Brown, 2005) | T6, F6, C6 |
| Mean kurtosis | $\sqrt{\frac{\sum_i (i-\bar{i})^4 . E_i}{E.RMS_i^4}}$ | (Tucker and Brown, 2005) | **T7**, F7, C7 |
| Shannon entropy [a] | $-\sum_j p(s_j).\log_2\left(p(s_j)\right)$ | (Esmaili *et al.*, 2004; Han *et al.*, 2011) | T, F, C 8 to 10 (**F8**) |
| [a] Bin numbers for probability estimation: 5, 30, 500 | | | |
| Rényi 'entropy' [b] | $\frac{1}{1-\alpha}\cdot\log_2\left(\sum_j p(s_j)^\alpha\right)$ | (Han *et al.*, 2011) | T, F, C11 to 12 (**F11, F12, C12** |
| [b] Bin numbers for probability estimation: 30, $\alpha = 2$ and $\infty$ | | | |
| Rate of attack | $\max_i\left(\frac{s[i]-s[i-1]}{n}\right)$ | (Tucker and Brown, 2005) | T13, F13, C13 |
| Rate of decay | $\min_i\left(\frac{s[i]-s[i+1]}{n}\right)$ | (Tucker and Brown, 2005) | T14, F14, C14 |
| Threshold crossing rate [c] | $\frac{\#(\text{Threshold Crossing})}{n}$ | (Fagerlund, 2007; Huang *et al.*, 2009) | T, F, C 15 to 18 (**T15, T16**) |
| Silence ratio [c] | $\frac{\#(\text{s where s<threshold})}{n}$ | (Mitrovic *et al.*, 2006) | T, F, C 19 to 22 (**F22, F21, F20**) |
| [c] Signal maximum normalized to 1 and different threshold values: 0.2, 0.4, 0.6 and 0.8 | | | |
| Mean | $\mu = \frac{\sum_i s[i]}{n}$ | | T23, **F23**, C23 |
| Max over mean | $\frac{\max_i s[i]}{\mu}$ | | T24, F24, C24 |
| Min over mean | $\frac{\min_i s[i]}{\mu}$ | | T25, **F25**, C25 |
| Energy measurements[d] | Energy standard deviation, energy skewness, energy kurtosis | (Foote, 1997) | T, F, C 26 to 28 **F28, T26, F17, F26** |
| [d] Energy measurements are features computed from the signal energy $E(t) = x(t)^2$ rather than on the signal itself. | | | |

2002). `Impulsions` might also be linked to feeding activities according to (Amorim *et al.*, 2004). `Drums` and `Roars` could also be related to courtship behavior (Mann *et al.*, 2008). Additionally, it is worth noting that the fish sounds listed in the biology literature are generic: the four classes identified here are not specific to the data used in this study (Amorim, 2006). As a consequence, the architecture we propose to automatically classify fish sounds can be used for recordings other than those used in this study.

### C. Dataset preprocessing

A dataset of observations is needed to train and test any classification model, with the dataset construction here. A labeled dataset is a database of signals in which each observation has been assigned to its corresponding semantic class detailed here. For our application, this meant considering a large number of fish sounds of each of the four considered classes. To distinguish fish sounds from uninteresting sounds, we also considered a fifth class of background noise. Once built, the dataset was used to train and test the classification model: the model learns to distinguish and recognize the various classes from the

| Ref. | Area | Depth | Duration |
|---|---|---|---|
| 1 | Healthy sea-grass meadow | -20m | 3.5 days |
| 2 | Healthy sea-grass meadow | -12m | 1 day |
| 3 | Lower sea-grass meadow / sand border | -38m | 1 day |
| 4 | Damaged meadow | -12m | 1 day |
| 5 | Rock | -12m | 1 day |

TABLE II. Description of the area-specific recordings used in the present study.

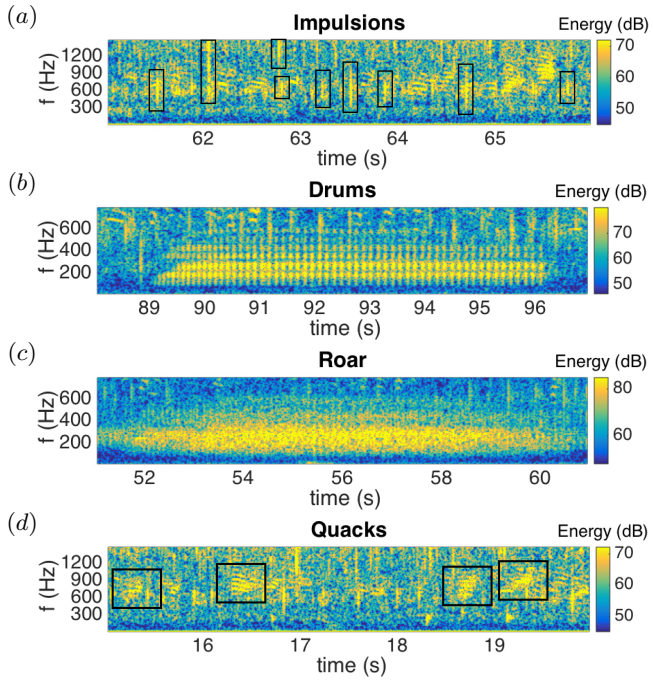


FIG. 1. Description of the four classes, and their corresponding spectrograms.
(a) Impulsions. Length: 2 ms to 20 ms. Presence: $< 5/s$. Spectrogram with Gaussian windows on 8192 points.
(b) Drums. Length: 5 s to 20 s. Presence $< 5/\text{min}$. Spectrogram with Gaussian windows on 16384 points.
(c) Roars. Length: 10 s to 25 s. Presence: $< 2/\text{min}$. Spectrogram with Gaussian windows on 16384 points.
(d) Quacks. Length: 150 ms to 300 ms. Presence: $< 5/s$. Spectrogram with Gaussian windows on 8192 points.

observations of the database. This implies that, ideally, the dataset should contain all of the variability of the phenomenon under study. As this is physically impossible, the idea was to gather as many observations as possible, to characterize a given class as completely as possible. The dataset used to build the model is directly linked to the model capabilities. Without data covering a wide spectrum of observations, it is very difficult for the model to analyze newly recorded data. It is therefore necessary to consider a *large* dataset that covers the range of the phenomena under study. Further explanations on supervised machine learning are given in Section I B 1.

For the present study, 913 observations were manually identified from the underwater recordings by an expert: 91 Impulsions, 114 Drums, 36 Roars, 205 Quacks, and 467 Background. We hereafter detail the process. All of these observations were extracted from continuous labeling of 10 min at the sand / sea-grass interface (Table II, area reference #3). This particular area was selected because it appeared to host the most varied recordings. The labeled period was recorded on August 5, 2014, at 10 pm, and was selected as a particularly rich recording (i.e., gathering many fish sounds). The recordings were continuously labeled using a sliding window of fixed length $\Delta_t = 0.5s$ and two bandwidths. We chose to focus our analysis on the frequency ranges of 50 Hz to 450 Hz, and 400 Hz to 900 Hz, as most of the fish sounds in the recording were in these frequency bands. The original recordings were previously down-sampled to $f_s = 10400Hz$. Each observation therefore had a fixed length of $\Delta_t = 0.5s$, and belonged to one of the two frequency ranges that were analyzed. The use of a sliding window of fixed size led to some calls being considered as various observations; e.g., `Drums` and `Roars` are long calls (10-30 s), and were therefore separated into several consecutive observations. The $0.5s$ window length was empirically determined as the minimum duration needed to distinguish the five classes. This was longer than a single `Impulsion` or a single `Quack`, and shorter than either a full `Drum` or a full `Roar` call. However, a minimum of $0.5s$ is needed to identify a `Drum` or a `Roar` as such. Alternatively, and depending on the data, the continuous analysis proposed in this study can be carried out on windows of different sizes and in other frequency bands. Any observation where the class was not clearly identified by the experts belonged to the `Unknown` class, and were disregarded for the learning stage (but not for the testing; see Section IV B). Alternatively, observations that contained no fish sounds and no unidentified sound were labeled as `Background`. The labeling step is illustrated in Figure 2: each observation of the dataset is a signal of length 0.5 s and filtered in its bandwidth. They can be visualized as an extract of the spectrogram.

## IV. RESULTS

In this Section, we present the various experiments that were conducted, along with their goals and results.

**Model validation:** in which we show the performance of the model using cross-validation in order to validate the proposed architecture.

**Features selection:** in which the features influence is stressed, and where we provide an illustration of the feature selection process.

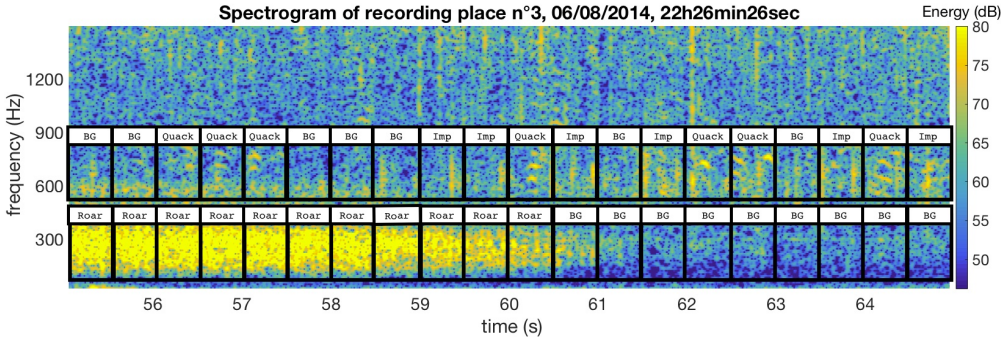


FIG. 2. Illustration of the annotation process. All of the black boxes have the same width, corresponding to $0.5s$, and the same height corresponding to the frequency range ($f_{max} - f_{min} = 400$Hz in this study). An observation is a $0.5s$ portion of the recording, filtered between its $f_{min}$ and $f_{max}$. The spectrogram was generated using a sliding Gaussian window of 16384 points.

in which we use the proposed architecture to process continuous recordings of underwater acoustic data. In particular, we use the labeled data recorded at the lower sea-grass border with the sand to analyze the data recorded at the same location but latter (Table II, area Ref3), and the data recorded with healthy sea-grass (Table II, area Ref2). The validation process is carried out for 10 min for each area.

The code used for this study and in this project was written in Python using the scikit-learn library and run on a laptop computer.[2]

### A. Model validation

In this first part of the study, the proposed architecture is tested using cross-validation. A total of 913 observations belonging to five classes are considered: the four positive classes associated to fish sounds (`Impulsions`, `Drums`, `Roars` and `Quacks`), and a generic `Background` class. Cross-validation with $\alpha = 0.7$ is used to determine the best values for the hyper-parameters of the learning algorithms, and to validate the model performances. When using SVM, the best accuracy of the data is obtained with a Gaussian radial basis function kernel of parameter $\gamma = 2^{-7}$, and with a cost parameter $C_{SVM} = 512$. These values are obtained after a grid search for the optimum values, and then they are kept constant. The accuracy of the results varies smoothly through the grid. For RF, the number of trees is fixed to 200, as a compromise between performance and computation time. The entropy is used as impurity measurement since it leads to better results than Gini index, $\sqrt{d}$ features are considered at each node with $d$ the total number of features, and the trees are not pruned. The overall accuracy reaches $95.3\% \pm 0.76\%$ when using all of the features and RF as classifier, and $95.0\% \pm 0.88\%$ when using SVM. These results validate both the architecture and the features used. Two main conclusions are drawn from those numbers: first, the overall proposed process to automatically classify the fish sounds is validated. Second the learning algorithm has, as was expected, a limited influence on the results.

### B. Features selection

In this second experiment, we study the influence of the different features. In particular, we show that the features have a greater impact on the results than the choice of the learning algorithm. The accuracy when comparing the influence of the feature sets (`Time`, `Frequency`, `Cepstral` and `All`) and the learning algorithms (SVM, RF) are given in Table III. The influence of the feature sets is of particular interest here. When comparing the accuracy of the results with `Time`: $90.1 \pm 2.0\%$; `Frequency`: $90.1 \pm 2.7\%$; and `Cepstral` features: $91.4 \pm 3.0\%$, it is interesting to note that the feature set influence is no so important in this case. Those accuracy values correspond to RF but SVM leads to the same conclusion: each domain contains enough discriminative content, but combining the three leads to better performances. This phenomenon would suggest that the discriminative information needed for the classification is spread in the various domains. Once again, numbers show that results are steady regarding the learning algorithm that is used.

We therefore study the feature selection issue, and as explained in Section II, we use RF features weight to select the most important features. More specifically, Figure 3 (b) shows the individual importance of the features and Figure 3 (a) displays the mean evolution accuracy when the dimension $d$ of the feature vector increases: from the most important feature, to the two most important, and so on. Analysis of the features weights leads us to build two subsets of features of decreasing importance: the `most valuable features (MVF)` (Figure 3, red dots) and the `valuable features (VF)` (Figure 3, blue dots). The `MVF` and `VF` are selected as the minimum features subsets leading to stable results: using more features does not significantly increase the performance of the classification system. The accuracy when using `MVF` and `VF` are also reported in Table III. The `MVF` contains only three features: the energy kurtosis computed from the spectrum (`F28`), the mean kurtosis computed from the waveform (`T7`), and the threshold crossing rate also computed from the waveform (`T15`). Considering the mean accuracy over the five classes when using the `MVF`, it reaches 91.5% and 91.3% for RF and SVM, respectively. This result is essential for real-time applications and embedded systems with limited computational costs and storage capabilities. If we consider the 19 first features with `VF` (including features from `MVF`), the global accuracy reaches 95.6% and 94.7% for RF and SVM, respectively. Similar numbers are obtained when using the feature set `All`. This result shows that all of the features are actually not needed to obtain good classification results, and it also reflects on correlations between some of the features. Furthermore, it is worth noting that features of `VF` contain descriptors computed on the signals represented in the time, frequency, and cepstral domains, encouraging to consider several representations of each observation.

The impact of the most important features on the class by class accuracy is also particularly relevant and is reported in Table IV. In particular, it reveals that valuable features are not equally important depending on the considered class. The second most important feature for example (`T7`, mean kurtosis computed from the time domain) has a great impact on `Background`, `Drums` and `Roars`, has no effect on `Impulse`, but is detrimental to `Quacks` since their accuracy drops from 59.3% to 57.8% when using a second feature to represent the observations.

TABLE III. General Results for the Automatic Classification of Fish Sounds. Accuracy results are compared depending on (i) the feature set used (time, frequency, cepstral, all or MFCC) and (ii) the learning algorithm (RF or SVM). Subsets of the most important features are also considered: `MVF` and `VF`. Learning rate $\alpha = 0.7$.

| Feature set | Dimension $d$ | Accuracy | |
|---|---|---|---|
| | | RF | SVM |
| `All` | 84 | $96.9 \pm 2.0\%$ | $96.5 \pm 1.6\%$ |
| `Time` | 28 | $90.1 \pm 2.0\%$ | $91.2 \pm 1.8\%$ |
| `Frequency` | 28 | $91.1 \pm 2.7\%$ | $90.7 \pm 3.1\%$ |
| `Cepstral` | 28 | $91.4 \pm 3.0\%$ | $90.8 \pm 2.7\%$ |
| `MVF`[a] | 3 | $91.5 \pm 0.85\%$ | $91.3 \pm 0.82\%$ |
| `VF`[b] | 19 | $95.6 \pm 0.79\%$ | $94.7 \pm 0.82\%$ |
| `MFCC` | 26 | $72.5 \pm 3.3\%$ | $70.0 \pm 6.0\%$ |

[a] with `MVF` for Most Valuable Features
[b] with `VF` for Valuable Features

TABLE IV. Class by class accuracy for feature vectors made of the 1st to 5th most important features, according to Random Forest features weights. Features are designated by their references, as specified in Table I.

| | Accuracy (%) | | | | |
|---|---|---|---|---|---|
| Background | 72.1 | 76.0 | 77.2 | 78.6 | 84.2 |
| Impulse | 63.5 | 63.8 | 67.3 | 68.5 | 72.0 |
| Drums | 62.8 | 91.1 | 91.7 | 91.8 | 93.8 |
| Quacks | 59.3 | 57.8 | 58.2 | 61.3 | 65.8 |
| Roars | 67.9 | 92.9 | 94.3 | 93.5 | 95.0 |
| New feature ref. | F28 | T7 | T15 | F11 | F1 |

### C. Continuous analysis of underwater acoustic recordings

Finally, this section reports on the use of our proposed method to automatically analyze continuous underwater recordings. More particularly, we train a model on the dataset presented in Section III C that contains 91 `Impulsions`, 114 `Drums`, 36 `Roars`, 205 `Quacks` and 467 `Background` observations. These observations are extracted from a continuous labeling of 10 min of recording on August 5, 2014, at 10 pm. The model is trained with SVM and the `All` features. The model is then used in an applicative context to analyze continuous recordings in two different configurations. The first one tests the model performance on continuous recordings: the test signals were recorded on August 5, 2014, between 10:27 pm and 10:37 pm, that is half an hour after the acquisitions used for training the model. The two data sets (i.e., the one used for training and the one for test) were recorded on the same area. The second test configuration considers a set of recordings that was randomly selected among the other recording areas. These were registered in Area Ref #2 given in Table II (sea grass at 22 m in depth, compared to sea grass/sand interface at a depth of 38 m, for the learning observations) on August 6, 2014, at 11 pm. In both configurations, the recordings are continuously processed in the two bandwidths on which this study is focused (i.e., 50-450 Hz and 500-900 Hz), with a sliding window of duration 0.5 s. To reduce the computational burden in both the training and validation processes, no overlap is considered between consecutive windows. The threshold value is empirically fixed to $0.8 \in [0, 1]$. Classification results are presented in Tables V and VI.
In this configuration, class by class accuracy and precision results are presented, in order to better explicit the false alarm rates. The first configuration was decided to test the model on continuous signals that were recorded at a different time than the learning observations, in order to avoid similarities between the signals. The confusion matrix of this test is presented in Table V and several conclusions can be drawn. First, very few errors are noticed. Regarding `Drums`, 133 observations are correctly detected, and two mislabeled in the `Background` class. Similarly, 208 `Impulsions` are correctly detected, two observations are confused for `Background` and one for `Drums`. Comparable results are obtained for `Quacks` and `Roars`, with 245 and 58 correct detections for no and two errors, respectively. The associated accuracy is therefore extremely high going up to 95.7% The study of the `Unknown` observations is also particularly relevant

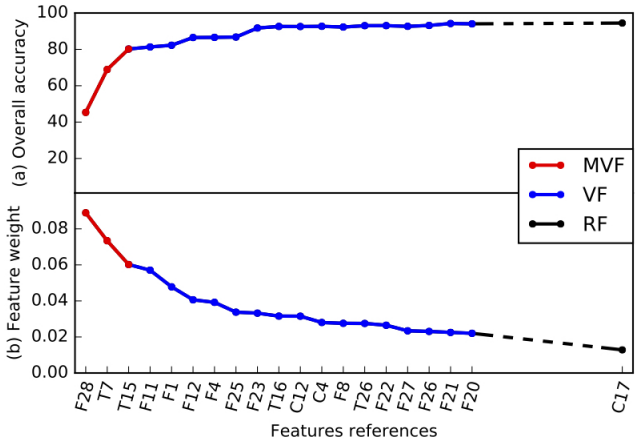


FIG. 3. Evolution of (a) mean accuracy when using feature vectors of increasing dimension $d$. Features individual weights are shown in (b). Features are references as indicated in Table I with `T` for Time, `F` for Frequency and `C` for Cepstral domain. For instance, `F28` refers to the energy kurtosis computed from the spectral domain, `T7` to the mean kurtosis computed from the time domain, and `T15` to the threshold crossing rate ($t = 0.2$) computed from the time domain. In red, the Most Valuable Features (`MVF`), in blue the Valuable Features (`VF`) and in black the regular features (`RF`). The valuable features are highlighted in bold font in Table I.

since we observe 743 observations for which the model does not reach any decisions. Among those, 624 are also not identified by experts, either because the observation is too noisy to be identified, or because it contains more than one class, or because the sound does not match any of the four positive classes. Between 11 (`Impulsions`) and 39 (`Background`) observations per class should have been attributed to a class, suggesting that the probability threshold could be decreased, and adjusted to each class. Precision results also lead to particularly valuable information, since they are systematically above 97% for the four fish sound classes: the false alarm rate is extremely low, which clearly validates the use of the model in real conditions. Results from the experiment are particularly conclusive since the overall high performance and the precision prove the interest of such tools for analyzing continuous data.

The second set of observations are considered to analyze the model generalization capacities since the model is tested on recordings from a different underwater area. Results of this experiment are presented in the confusion matrix in Table VI and very similar conclusions can be drawn from the detected observations. Almost all detected observations are correctly assigned to their classes. Depending on the class, a maximum of 25 errors out of 885 correct detections are found for the `Background` class, and regarding positive classes, 6 errors for 46 detections and 4 errors for 25 detections are made respectively for `Impulsions` and `Drums`. It is also interesting that no `Roars` are detected, which is confirmed by the analysis done by experts. Very good detection results are therefore achieved, even when the model is tested in a different area compared to the learning observations. It is relevant to notice that in this case, the number of rejected observations (`Unknown` class) is greater: 1024 compared to 743 in the previous configuration. Out of those observations, 651 are also rejected by the experts, but the others are missed by the model. Generally speaking, accuracy and precision rates are lower than when the training and testing recording places are similar, but still relatively high. A conclusion on those results is the need to lower the threshold in this configuration: when the learning and testing areas are different, test observations are less likely to be similar to the ones used in the training, and the probabilistic outputs of the model are therefore lower than in the previous configuration. Those results recommend the use of such a method for the continuous and real-time analysis of underwater recordings. In particular, large datasets can be automatically analyzed and conclusions can be drawn regarding the fish populations, their evolution in time, and their movement from one area to another.

As for the computation times, each set of recordings (i.e., duration of ten minutes) was analyzed in about 4 min on a laptop computer, thereby validating the use of this method for real-time applications. As a reminder, `all` features were used for this analysis, and thus the computation times could be decreased if only selected features are used.

TABLE V. The five considered classes are Background (`B`), Roars (`R`), Drums (`D`), Quacks (`Q`) and Impulses (`I`). A sixth class for rejected observations is considered and referred as Unknown (`U`). The average accuracy reaches: 93.4%. Testing observations recorded at a different time than learning observations.

| | | True Class (ground truth) | | | | | | Precision |
|---|---|---|---|---|---|---|---|---|
| | | B | D | I | Q | R | U | |
| | B | **969** | 2 | 2 | | 2 | 23 | 97.1% |
| | D | | **133** | 1 | | | 2 | 97.8% |
| Predicted | I | | | **208** | | | | 100% |
| Class | Q | 5 | | | **245** | | 2 | 97.2% |
| | R | | | | | **58** | 1 | 98.3% |
| | U | 39 | 16 | 11 | 40 | 13 | **624** | 84.0% |
| Accuracy | | 95.7% | 88.1% | 93.7% | 86.0% | 79.5% | 95.7% | |

TABLE VI. The five considered classes are Background (`B`), Roars (`R`), Drums (`D`), Quacks (`Q`) and Impulses (`I`). A sixth class for rejected observations is considered and referred as Unknown (`U`). The average accuracy reaches: 80.9%. Testing observations recorded at a different time and place than learning observations.

| | | True Class (ground truth) | | | | | | Precision |
|---|---|---|---|---|---|---|---|---|
| | | B | D | I | Q | R | U | |
| | B | **885** | 6 | 4 | | | 21 | 96.6% |
| | D | 3 | **40** | | | | 8 | 78.4% |
| Predicted | I | 16 | | **21** | | | | 56.8% |
| Class | Q | 6 | | | **256** | | | 97.7% |
| | R | | | | | | | |
| | U | 207 | 24 | 15 | 127 | | **651** | 63.6% |
| Accuracy | | 79.2% | 57.1% | 52.5% | 66.8% | | 95.7% | |

## V. DISCUSSION

We discuss here several issues around the fish sounds classification system and its limitations.

In particular, it is relevant to discuss the limitations of the propose approach in order to better evaluate the proposed results. The current main limitation of the method is the use of a fixed window for the analysis: some observations contain more than one class. Typically, `Quacks` and `Impulsions` can sometimes be found within the same window. The model then recognizes

properties of both classes and output probabilities are split between the main classes. However, both are often lower than the threshold that was fixed, leading to their rejection in terms of classification. To overcome this limitation, the study could be carried out on temporal windows of various length: smaller windows would be less likely to contain more than one class. The use of a sliding window also prevents temporal coherence in the model; e.g., for call counting operations, a temporal regularization would help to identify complete calls from several detected observations. This is particularly relevant for the long calls, such as `Roars` and `Drums`, which are detected as a succession of windows including the same class. To address this limitation, the use of hidden Markov models is currently under consideration, since they are effective in implementing a temporal regularization of the classification results.

Another limitation of the current system for the analysis of continuous recordings is the need to use a threshold. If a threshold is not used, all of the windows are classified between the four positive classes and `Background`. However, according to the interpretation done by experts, some of the windows are `'Unknown'`: if the fish are far away and the effects of the propagation are non-negligible, if different classes occur in the same window (sometimes up to three), or if the sound does not fit in any of the classes (unknown sounds), it is not possible for the experts to classify these observations. It is therefore necessary to threshold the output prediction probabilities to reject such observations. The thresholding operation, however, raises the issue of the threshold choice: if it is too high, only well-defined observations will be detected, and many will be missed; if it is too low, many observations that are `Unknown` for the experts will be forced into a class. Ideally, a different threshold should be decided upon for each class. A promising development of the existing model would be to perform an analysis of the `'Unknown'` observations in order to detect new classes of sounds. For example, classes related to anthrophony or geophony can be considered, and we can in particular think of boat, rain, or thunder sounds. The use of an unsupervised approach might also be envisaged, as the main limitation of supervised learning is the need for a labeled dataset. In many cases, including in this study, the technical difficulty to label the data is real and highly time consuming if done manually, which is a limiting factor. Unsupervised learning might also be interesting for a study of the observation variability within a considered class: as explained in Section III B, some classes have intra-class specifies that will be related to species or the identification of an individual; e.g., `Drums` are emitted with various pulse frequencies, while `Impulsions` are emitted across a heterogeneous frequency range.

### A. Comparison to the state of the art

As the state of the art in automatic classification of fish sounds is relatively limited, we compare the proposed method in terms of features to the use of `MFCCs` as descriptors. As explained in Section I B, `MFCCs` were originally designed for speech processing purposes, but have since been used in many applications related to automatic classification of transient signals. Comparing `MFCCs` to the `All` features, we obtain accuracies when using `MFCCs` of $72.5 \pm 3.3\%$ for RF and $70.0 \pm 6.0\%$ for SVM, while the `All` features reach 95.3% for RF and 95.0% for SVM. The feature set we propose for this application thus leads to more accurate results, and is actually significantly better adjusted. Indeed, `MFCCs` were originally developed to represent speech data, which are particularly different from the data used in the present study, in terms of their frequency range, shape, and other details. We thereby stress that the features proposed here are generic, and can also be used to represent more general transient signals. The same conclusion can be drawn when comparing the `MFCCs` with the `Cepstral` features, where the accuracy reaches $91.4 \pm 3.0\%$ for RF and 90.8 $pm2.7\%$ for SVM. The comparison between these two feature sets stands as they both describe the Cepstral domain. However, the feature set we propose here performs better than the `MFCCs`, once again stressing the importance of the feature choice and validating the proposed features. One reason for this might be that the `MFCC` is an ordinate representation, while the `Cepstral` feature set is not. All of the data related to the use of `MFCCs` features are presented in Table III, and all 26 `MFCC` coefficients were considered in this study.

## VI. CONCLUSION AND PROSPECTS

This study addressed the monitoring of fish populations by focusing on their emitted sounds since they can be associated to their activities. Specifically, this study aimed at building a supervised automatic fish sound classifier. The proposed classification system exceeds 95% of correct classification rate for five different classes (four fish sounds and background noise). The key points of the proposed architecture are: (i) the comprehensive set of features that was defined for describing the acquired signals; (ii) the feature extraction process, which proposes to extract features from three different representations of observations (time, spectral, cepstral); and (iii) the uniqueness of the labeled dataset presented. The method we propose is also used to process continuous recordings with excellent results. Such tools can therefore be used to analyze large datasets and perform real-time analysis of underwater recordings. The consequences for the monitoring of such areas are particularly significant.

The currently ongoing prospects of this study include the analysis of several days of recordings to draw conclusions regarding the fish populations present in the monitored area. This analysis will be conducted using the proposed system. For the technique of the method and its improvement, the use of a more comprehensive dataset that includes geophonic and anthrophonic events, for instance, is under consideration. The use of unsupervised methods to analyze the observations rejected by the system (e.g., `Unknown` class) is also under investigation. Finally, the use of a hidden Markov model is under

consideration, to exploit the temporal coherence of the recordings and to deal with fish calls of various lengths.

## ACKNOWLEDGMENTS

This study was supported by a grant from Labex OSUG@2020 (Investissements d'avenir - ANR10 LABX56) and DGA/MRIS Geosciences. GIPSA-Lab SIGMAPHY is part of Labex OSUG@2020 (ANR10 LABX56).

[1]In this study, as in many machine-learning problems, data is referred to as *data*, *observations*, *examples*, or acoustic *signatures*.

[2]The code for the automatic and continuous processes will shortly be available on GitHub.

Acevedo, M. a., Corrada-Bravo, C. J., Corrada-Bravo, H., Villanueva-Rivera, L. J., and Aide, T. M. (**2009**). "Automated classification of bird and amphibian calls using machine learning: A comparison of methods," Ecological Informatics **4**(4), 206–214.

Amorim, M., Stratoudakis, Y., and Hawkins, a. D. (**2004**). "Sound production during competitive feeding in the grey gurnard," Journal of Fish Biology **65**(1), 182–194.

Amorim, M. C. P. (**2006**). "Diversity of sound production in fish," Communication in fishes **1**, 71–104.

Bedoya, C., Isaza, C., Daza, J. M., and López, J. D. (**2014**). "Automatic recognition of anuran species based on syllable identification," Ecological Informatics **24**, 200–209.

Bellman, R. (**1956**). "Dynamic Programming and Lagrange Multipliers.," Proceedings of the National Academy of Sciences of the United States of America **42**(10), 767–769.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (**1992**). "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the fifth annual workshop on Computational learning*, pp. 144–152.

Breiman, L. (**2001**). "Random forests," Machine learning 5–32.

Chen, W. P., Chen, S. S., Lin, C. C., Chen, Y. Z., and Lin, W. C. (**2012**). "Automatic recognition of frog calls using a multi-stage average spectrum," Computers and Mathematics with Applications **64**(5), 1270–1281.

Chesmore, E. D. (**2001**). "Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals," Applied Acoustics **62**(12), 1359–1374.

Clemins, P. J., and Johnson, M. T. (**2006**). "Generalized perceptual linear prediction features for animal vocalization analysis.," The Journal of the Acoustical Society of America **120**(1), 527–534.

Costanza, R., d'Arge, R., De Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'neill, R. V., Paruelo, J. *et al.* (**1997**). "The value of the world's ecosystem services and natural capital," nature **387**(6630), 253–260.

Couvreur, C., Fontaine, V., Gaunard, P., and Mubikangiey, C. G. (**1998**). "Automatic classification of environmental noise events by hidden Markov models," Applied Acoustics **54**(3), 187–206.

Dennis, J., Tran, H. D., and Li, H. (**2011**). "Spectrogram image feature for sound event classification in mismatched conditions," IEEE Signal Processing Letters **18**(2), 130–133.

Deshpande, H., Singh, R., and Nam, U. (**2001**). "Classification of music signals in the visual domain," in *Proceedings of the COST-G6 Conference on Digital Audio Effects*, pp. 1–4.

Dong, X., Towsey, M., Zhang, J., Banks, J., and Roe, P. (**2013**). "A novel representation of bioacoustic events for content-based search in field audio data," in *Digital Image Computing: Techniques and Applications (DICTA), 2013 International Conference on*, IEEE, pp. 1–6.

Dos Santos, M. E., Modesto, T., Matos, R. J., Grober, M. S., Oliveira, R. F., and Canario, A. (**2000**). "Sound production by the lusitanian toad fish, halobatrachus didactylus," Bioacoustics **10**(4), 309–321.

Duda, R. O., Hart, P. E., and Stork, D. G. (**2001**). *Pattern Classification* (John Wiley & Sons).

Eronen, A., and Klapuri, A. (**2000**). "Musical instrument recognition using cepstral coefficients and temporal features," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, IEEE, Vol. 2, pp. II753–II756.

Esfahanian, M., Zhuang, H., and Erdol, N. (**2014**). "Sparse representation for classification of dolphin whistles by type.," The Journal of the Acoustical Society of America **136**.

Esmaili, S., Krishnan, S., and Raahemifar, K. (**2004**). "Content based audio classification and retrieval using joint time-frequency analysis," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 5, pp. 9–12.

Fagerlund, S. (**2007**). "Bird Species Recognition Using Support Vector Machines," EURASIP Journal on Advances in Signal Processing **2007**.

Foote, J. (**1997**). "A similarity measure for automatic audio classification," in *Proc. AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*.

Friedman, J., Hastie, T., and Tibshirani, R. (**2001**). *The elements of statistical learning*, **1** (Springer series in statistics New York).

Fujinaga, I., and MacMillan, K. (**2000**). "Realtime recognition of orchestral instruments," in *International Computer Music Association*.

Guo, G., and Li, S. Z. (**2003**). "Content-based audio classification and retrieval by support vector machines.," IEEE Transactions on Neural Network **14**(1), 209–215.

Han, N. C., Muniandy, S. V., and Dayou, J. (**2011**). "Acoustic classification of Australian anurans based on hybrid spectral-entropy approach," Applied Acoustics **72**(9), 639–645.

Heck, K. J. (**2003**). "Critical evaluation of nursery hypothesis for seagrasses," Marine Ecology Progress Series **253**, 123–136.

Huang, C.-J., Yang, Y.-J., Yang, D.-X., and Chen, Y.-J. (**2009**). "Frog classification using machine learning techniques," Expert Systems with Applications **36**(2), 3737–3743.

Langley, P. A. T., Flamingo, L., and Edu, S. (**1994**). "Selection of Relevant Features in Machine Learning," 127–131.

Lee, C. H., Chou, C. H., Han, C. C., and Huang, R. Z. (**2006**). "Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis," Pattern Recognition Letters **27**(2), 93–101.

Lim, T., Bae, K., Hwang, C., and Lee, H. (**2007**). "Classification of underwater transient signals using mfcc feature vector," in *Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on*, IEEE, pp. 1–4.

Lossent, J., Gervaise, C., and Iorio, L. D. (**2015**). "Cartographie de la biophonie des écosystèmes côtiers," 1–5.

Malfante, M., Dalla Mura, M., Mars, J. I., Metaxian, J. P., Macedo, O., and Inza, A. (**2017**). "Machine Learning for Volcano-seismic Signals: Challenges and Perspectives" .

Mann, D. A., Hawkins, A. D., and Jech, J. M. (**2008**). *Active and passive acoustics to locate and study fish*, **32** (Springer), p. 279.

Márquez-Molina, M., Sánchez-Fernández, L. P., Suárez-Guerra, S., and Sánchez-Pérez, L. A. (**2014**). "Aircraft take-off noises classification based on human auditory's matched features extraction," Applied Acoustics **84**, 83–90.

McCulloch, W. S., and Pitts, W. (**1943**). "A logical calculus of the ideas immanent in nervous activity," The Bulletin of Mathematical Biophysics **5**(4), 115–133.

McIlraith, A. L., and Card, H. C. (**1997**). "Birdsong recognition using backpropagation and multivariate statistics," IEEE Transactions on Signal Processing **45**(11), 2740–2748.

Mitrovic, D., Zeppelzauer, M., and Breiteneder, C. (**2006**). "Discrimination and retrieval of animal sounds," in *Multi-Media Modelling Conference Proceedings, 2006 12th International*, IEEE, pp. 5–pp.

Noda, J. J., Travieso, C. M., and Sánchez-Rodríguez, D. (**2016**). "Automatic taxonomic classification of fish based on their acoustic signals," Applied Sciences **6**(12), 443.

Pace, F., Benard, F., Glotin, H., Adam, O., and White, P. (**2010**). "Subunit definition and analysis for humpback whale call classification," Applied Acoustics **71**(11), 1107–1112.

Parmentier, E., Vandewalle, P., Frédérich, B., and Fine, M. L. (**2006**). "Sound production in two species of damselfishes (Poma-

centridae): Plectroglyphidodon lacrymatus and Dascyllus aruanus," Journal of Fish Biology **69**(2), 491–503.

Sattar, F., Cullis-Suzuki, S., and Jin, F. (**2016**). "Acoustic analysis of big ocean data to monitor fish sounds," Ecological Informatics **34**, 102–107.

Thode, A. M., Kim, K. H., Blackwell, S. B., Greene Jr, C. R., Nations, C. S., McDonald, T. L., and Macrander, A. M. (**2012**). "Automated detection and localization of bowhead whale sounds in the presence of seismic airgun surveys," The Journal of the Acoustical Society of America **131**(5), 3726–3747.

Thorson, R. F., and Fine, M. L. (**2002**). "Crepuscular changes in emission rate and parameters of the boatwhistle advertisement call of the gulf toadfish, Opsanus beta," Environmental Biology of Fishes **63**(3), 321–331.

Tucker, S., and Brown, G. J. (**2005**). "Classification of transient sonar sounds using perceptually motivated features," IEEE Journal of Oceanic Engineering **30**(3), 588–600.

Tyagi, H., Hegde, R. M., Murthy, H. A., and Prabhakar, A. (**2006**). "Automatic identification of bird calls using spectral ensemble average voice prints," in *Signal Processing Conference, 2006 14th European*, IEEE, pp. 1–5.

Vieira, M., Fonseca, P. J., Amorim, M. C. P., and Teixeira, C. J. (**2015**). "Call recognition and individual identification of fish vocalizations based on automatic speech recognition: An example with the lusitanian toadfish," The Journal of the Acoustical Society of America **138**(6), 3941–3950.

Wang, S., and Zeng, X. (**2014**). "Robust underwater noise targets classification using auditory inspired timefrequency analysis," Applied Acoustics **78**, 68–76.

Wimmer, J., Towsey, M., Planitz, B., Roe, P., and Williamson, I. (**2010**). "Scaling acoustic data analysis through collaboration and automation," in *e-Science (e-Science), 2010 IEEE Sixth International Conference on*, IEEE, pp. 308–315.

Yu, G., and Slotine, J.-J. (**2009**). "Audio classification from time-frequency texture," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, IEEE, pp. 1677–1680.

Zaugg, S., Van Der Schaar, M., Houégnigan, L., Gervaise, C., and André, M. (**2010**). "Real-time acoustic classification of sperm whale clicks and shipping impulses from deep-sea observatories," Applied Acoustics **71**(11), 1011–1019.

Zheng, F., Zhang, G., and Song, Z. (**2001**). "Comparison of different implementations of mfcc," Journal of Computer Science and Technology **16**(6), 582–589.