Automatic detection of fish sounds based on multi-stage classification including logistic regression via adaptive feature weighting

Ryosuke Harakawa, Takahiro Ogawa, Miki Haseyama, and Tomonari Akamatsu

Citation: The Journal of the Acoustical Society of America **144**, 2709 (2018); doi: 10.1121/1.5067373

View online: https://doi.org/10.1121/1.5067373

View Table of Contents: http://asa.scitation.org/toc/jas/144/5

Published by the Acoustical Society of America

Automatic detection of fish sounds based on multi-stage classification including logistic regression via adaptive feature weighting

Ryosuke Harakawa,[a] Takahiro Ogawa, and Miki Haseyama
Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060-0814, Japan

Tomonari Akamatsu
National Research Institute of Fisheries Science, Fisheries Research Agency, Yokohama, Kanagawa 236-8648, Japan

(Received 10 November 2017; revised 7 October 2018; accepted 16 October 2018; published online 8 November 2018)

This paper presents a method for automatic detection of fish sounds in an underwater environment. There exist two difficulties: (i) features and classifiers that provide good detection results differ depending on the underwater environment and (ii) there are cases where a large amount of training data that is necessary for supervised machine learning cannot be prepared. A method presented in this paper (the proposed hybrid method) overcomes these difficulties as follows. First, novel logistic regression (NLR) is derived via adaptive feature weighting by focusing on the accuracy of classification results by multiple classifiers, support vector machine (SVM), and k-nearest neighbors (k-*NN*). Although there are cases where SVM or k-*NN* cannot work well due to divergence of useful features, NLR can produce complementary results. Second, the proposed hybrid method performs multi-stage classification with consideration of the accuracy of SVM, k-*NN*, and NLR. The multi-stage acquisition of reliable results works adaptively according to the underwater environment to reduce performance degradation due to diversity of useful classifiers even if abundant training data cannot be prepared. Experiments on underwater recordings including sounds of *Sciaenidae* such as silver croakers (*Pennahia argentata*) and blue drums (*Nibea mitsukurii*) show the effectiveness of the proposed hybrid method. © *2018 Acoustical Society of America*.

https://doi.org/10.1121/1.5067373

[ZHM] Pages: 2709–2718

I. INTRODUCTION

In fisheries science, it is important to identify marine species for purposes such as estimation of distribution and abundance.[1] For environmental conservation and sustainable use of marine bioresources, it is necessary to develop an identification method that does not have influence on the ecosystem.[2]

To meet this necessity, methods that utilize marine organism sounds are useful because of their non-invasive nature for the resource stock.[3–7] Research on automatic detection of sounds from marine mammals such as whales and dolphins has attracted much attention.[8–12] However, there have been few studies that target fish sounds despite the fact that fish biology is an important field of study.[1] As pioneer works, methods for detecting sounds produced by fish were proposed.[13,14] Diep *et al.*[13] proposed a method for detecting human audible signals during spawning acts of shad fish (*Alosa fallax*) on the basis of a Gaussian mixture model (GMM)[15]-based approach. Since this method uses terrestrial recordings on the shore, however, it does not perform well for underwater recordings. Vieira *et al.*[14] proposed a method using underwater acoustic data that enables individual recognition and sound type classification of the Lusitanian toadfish (*Halobatrachus didactylus*) via a hidden Markov model (HMM)[16]-based approach. In contrast, Matsuo *et al.*[17] proposed a method for automatic detection of fish sounds from underwater acoustic data. Their method enables detection of fish sounds by rule-based filtering; however, their method has the limitation of species-specific parameter tuning requiring manual inspections of the sound of target fish by humans.

In summary, the conventional methods mentioned above have some drawbacks. Existing methods[8–14,17] require exclusive acoustic features of target fish sounds. The methods do not work well for other species due to the highly diverse features of fish and environmental sounds. Most underwater recordings contain few target sounds in long duration recording of environmental sounds. Manual collection of many fish sounds is very time-consuming. It is difficult to prepare a large amount of training data, which is generally required by supervised machine learning techniques.

To overcome this difficulty, we present a supervised machine learning method (the proposed hybrid method) for automatic detection of fish sounds from underwater recordings. Even if a large amount of training data cannot be prepared in the situation in which features of fish and environmental sounds are highly diverse, the proposed hybrid method works well through multi-stage classification including logistic regression via adaptive feature weighting. First, based on the

[a] Electronic mail: harakawa@lmd.ist.hokudai.ac.jp

paper in Ref. 18, we realize novel logistic regression (NLR) that enables adaptive feature weighting by focusing on the accuracy of classification results by multiple classifiers, i.e., support vector machine (SVM)[19] and *k*-nearest neighbors (*k-NN*).[20] Although there are cases in which a single classifier, SVM or *k-NN*, cannot work well due to divergence of useful features, NLR combined with the two methods can produce complementary classification results. Here, we sequentially apply machine learning algorithms including SVM, *k-NN*, and NLR with consideration of the accuracy of each classifier for fish sound detection. The multi-stage acquisition of reliable classification results works adaptively according to the target underwater environment to reduce performance degradation due to diversity of useful classifiers even if a large amount of training data cannot be prepared. Experimental results for underwater recordings from the coast of Japan, including vocalizations of *Sciaenidae*, show that the proposed hybrid method enables successful detection of fish sounds.

II. METHODS

In this section, we first describe the method for collecting underwater acoustic data used in this study. Next, we provide an overview of the proposed hybrid method and its details.

A. Data collection and a rule-based filter

We used underwater acoustic data including sounds of *Sciaenidae*. These species produce sounds underwater.[21] Stationary observation was performed from the coast of Kashima, Japan (35° 54′7.04″N, 140° 44′2.45″E). We used datasets including underwater acoustic data recorded on 0:00–5:59 and 14:00–23:59 of July 10, 2015 as examples for testing the performance of the proposed hybrid method. It should be noted that the recorded data contained stereo signals that were stored in MP3 format (128 kbps) and resampled at 44.1 [kHz].

Points including fish sounds were clarified by a rule-based filter designed on the basis of the literature[22] that examines the characteristics of the target fish sounds. Detection results by this rule-based filter specially designed for croakers' sounds[22] are the reference to evaluate the proposed hybrid method. The details of the filter are shown below.

- Audio signals are loaded every 2 097 152 points.
- If the ratio of a power spectrum for 100–500 [Hz] is higher than a threshold, we judge that the signals contain ship and wave noise and remove them from the subsequent processing.
- For the remaining signals, we reload signals every 16 384 points, i.e., 371 [ms]. Note that this length is approximately that of the target fish sounds.
- If the ratio of a power spectrum of the reloaded signals for 400–800 [Hz] is higher than a threshold, we judge that the signals contain fish sounds and select them as targets for subsequent analysis.
- A finite impulse response (FIR) filter[23] is applied to the selected signals to extract signals for which frequency components are 400–800 [Hz].
- The envelop of the extracted signals is detected to obtain the cyclic pulse structure. Here, we select pulses for which intervals are 10 [ms] or more and detect peaks. Note that filter length for the envelop detection is set to 100 points.
- We reload signals for which the length is 400 [ms] around the point with maximum amplitude of the obtained signals. If the maximum amplitude is higher than a threshold, subsequent analysis is performed.
- We calculate the power spectrum for 0–22.1 [kHz] and check whether the maximum of the spectrum is within 400–800 [Hz] or not in order to verify that the peak of the spectrum corresponds to the target fish sounds. If the maximum of the spectrum is within 400–800 [Hz], we obtain the signals as targets for subsequent analysis.
- Since signals with less than five pulses may be contact noise, signals with five or more pulses are selected from the obtained signals as targets for the subsequent processing.
- Since the target fish sounds have approximately equivalent pulse intervals, signals with pulse intervals for which the standard deviation is 0.02 or less and average is 30 [ms] or less are selected. Note that the first two pulse intervals are ignored since there are cases in which the target fish sounds have long intervals.
- The selected signals are adopted as the target fish sounds.

In the experiment shown later, we evaluated the performance of the proposed hybrid method by using the obtained points as ground truths of fish sound detection.

B. Overview of the proposed hybrid method

Figure 1 shows an overview of the proposed hybrid method, which is a multi-stage classification framework that combines SVM, *k-NN*, and NLR. As shown in the figure, even if each classifier, i.e., SVM and *k-NN*, cannot work well, successful fish sound detection becomes feasible by sequentially using NLR that integrates them. Here, we judge whether SVM and *k-NN* work well or not by monitoring accuracy, i.e., coincidence ratio of classification results and ground truths by the rule-based filter explained above. In the proposed hybrid method, after screening of irrelevant segments, we extract audio features from the segments that are likely to include fish sounds (Sec. II C). Next, we construct classifiers, SVM and *k-NN*, by using the audio features and derive NLR on the basis of classification results by SVM and *k-NN*; then, fish sound detection is performed by multi-stage classification (Sec. II D). Although features and classifiers that produce good detection results differ depending on the underwater environment, the multi-stage processing with consideration of the accuracy of each classifier enables successful detection of fish sounds.

C. Pre-processing of multi-stage classification

First, we perform screening of irrelevant segments for detection of fish sounds in order to reduce computational cost and realize accurate detection. Specifically, for each channel, we divide the underwater acoustic data into segments for which the interval and overlap are T and Δ, respectively. As shown in Fig. 2, the power spectrum of underwater environmental

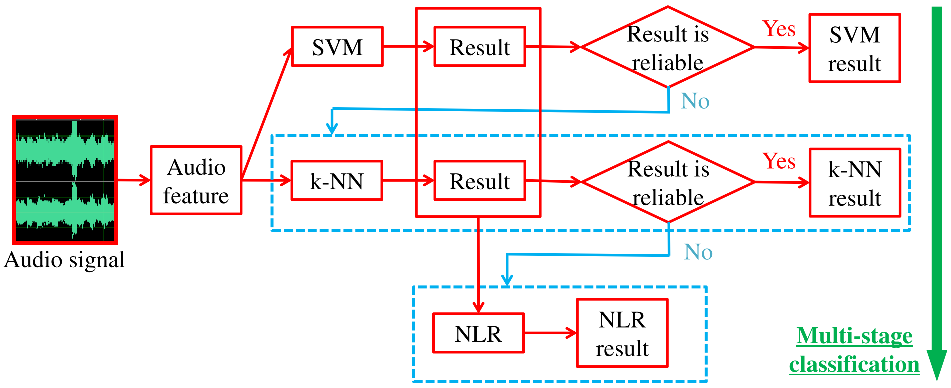
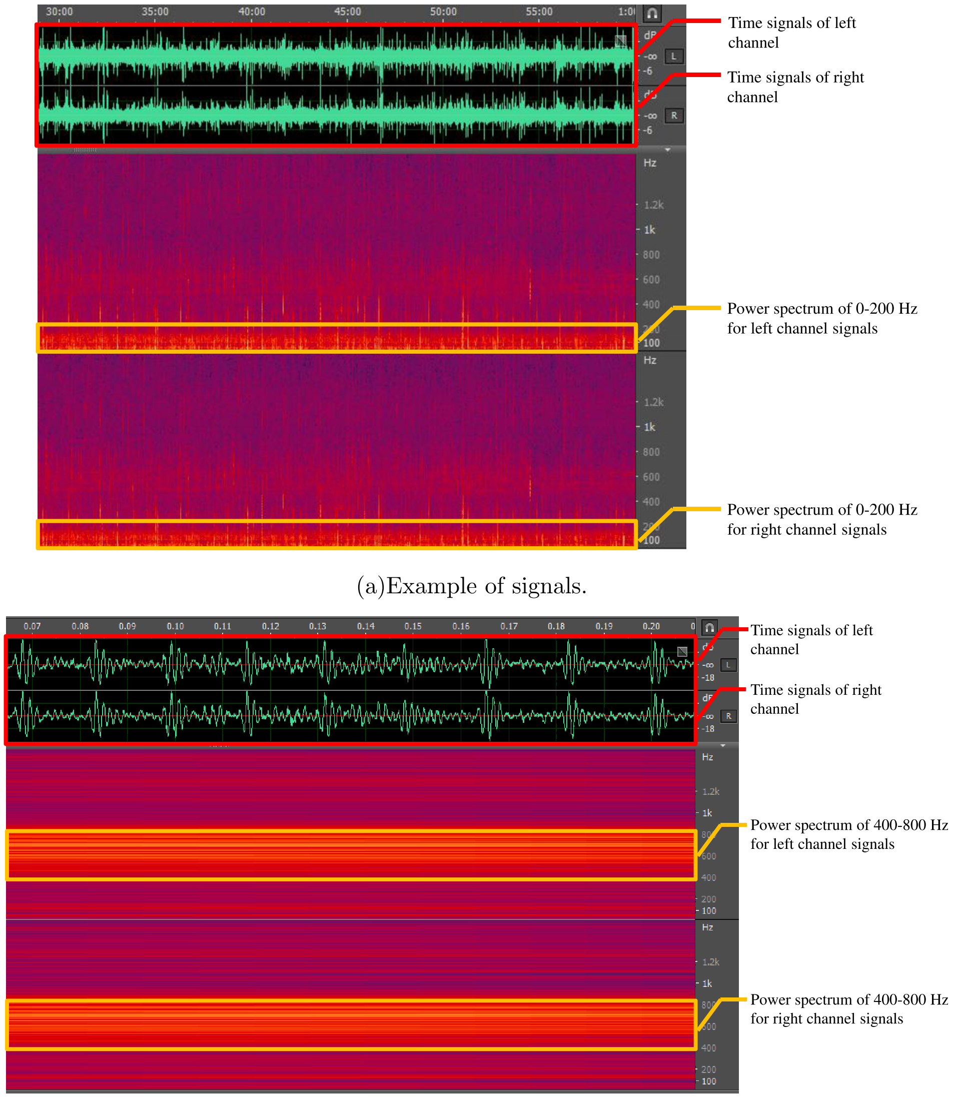


FIG. 1. (Color online) Overview of the proposed hybrid method. From audio signals, audio features are calculated after screening of irrelevant segments. SVM and *k-NN* are constructed by using the audio features and NLR is derived on the basis of classification results by SVM and *k-NN*. Fish sounds are detected via multi-stage classification, i.e., sequential application of SVM, *k-NN*, and NLR with consideration of the accuracy of each classifier.

sounds is densely distributed in less than 200 [Hz]. Since the environmental sounds are not necessary for detecting fish sounds, we apply a FIR filter[23] to each segment obtained from each channel. Thus, frequency components less than 200 [Hz] are removed from the segments. In the experiment shown later, we designed the FIR filter as a direct II transposed structure so that the filter order becomes 254 using the hamming window.

Second, we check whether the filtered segments may include fish sounds or not. For each channel, we calculate the sub-band energy ratio of 400–800 [Hz] components, i.e., the dominant frequency components (see Fig. 2), over 0–22050 [Hz] components. The frequency band focused on in this screening is the same as the reference data processing described previously.[22] However, since the paper[21] examined sounds of 20 kinds of fish and dominant frequencies of


30:00
35:00
40:00
45:00
50:00
55:00
Time signals of left channel
Time signals of right channel
Hz
1.2k
1k
800
600
400
100
Power spectrum of 0-200 Hz for left channel signals
Power spectrum of 0-200 Hz for right channel signals
0.07
0.08
0.09
0.10
0.11
0.12
0.13
0.14
0.15
0.16
0.17
0.18
0.19
0.20
Time signals of left channel
Time signals of right channel
Power spectrum of 400-800 Hz for left channel signals
Power spectrum of 400-800 Hz for right channel signals

(a)Example of signals.

(b)Detail of signals containing fish sounds.

FIG. 2. (Color online) Example of time signals and power spectrum of underwater acoustic data. (a) It can be seen that underwater environmental sounds are densely distributed in less than 200 [Hz]. (b) It can be seen that the target fish sounds are densely distributed in 400–800 [Hz].

most fish are less than 800 [Hz], the screening can also be applied to other species. If each sub-band energy ratio for both channels is less than Th_s, we judge that the segments do not include the target fish sounds to avoid further analysis. In this way, the proposed hybrid method enables screening of irrelevant segments from the enormous amount of underwater recordings to reduce computational cost and realize accurate detection. This first screening is similar to the rule-based filtering[22] described in Sec. II A; however, the proposed signal processing uses more general spectral and temporal features that can be applied to a wide variety of fish sounds.

After screening of irrelevant segments, we extract underwater acoustic features by using the remaining segments as follows. We reload segments to capture the whole length of the target fish sounds. In each remaining segment, we obtain monaural signals and compute a point where the amplitude is maximum. Then we reload segments from underwater acoustic data for which the length is T around the point with maximum amplitude. Furthermore, we apply the FIR and pre-emphasis filters[24] to each channel of the reloaded segments, respectively. Thus, we can extract signals for which frequency components are from 400 to 800 [Hz] and those for which low frequency components are removed. For the obtained signals, we calculate the feature vectors, which consist of the following elements:

- frequency centroid;
- frequency band width;
- spectral rolloff (85 percentile);
- zero-crossing rate;
- root-mean-square energy;
- mel-frequency cepstrum coefficients (MFCC);[25]
- polynomial features (these features are coefficients of fitting an N_pth order polynomial to the columns of a spectrogram. In the experiment, we varied N_p from 1 to 10);
- auto correlation (auto correlation whose dimension is D_a is calculated. In the experiment, D_a was defined as 100);
- root mean square (RMS) of amplitude;
- the number, maximum, minimum, mean, and standard deviation of peaks[26] (the sharp peaks of the obtained signals are detected via continuous wavelet transform[26]); and
- the maximum, minimum, mean, and standard deviation of amplitude for the obtained signals.

Here, the effectiveness of frequency centroid; frequency band width; spectral rolloff; zero-crossing rate; root-mean-square energy; and MFCC for audio classification is verified in the previous literature.[27–29] Furthermore, polynomial features, auto correlation, RMS of amplitude, statistics of peaks and signals are extracted for representing unique characteristics in the spectrum and shape of fish sounds.[21] In this way, we employ the features with consideration of the usefulness in machine learning and ecological fields. It should be noted that features for signals through the FIR and pre-emphasis filters are concatenated into one vector, and two kinds of feature vectors (two channels) are thus obtained per segment. In the subsequent phases, we judge whether the obtained segments include fish sounds or not by using the calculated feature vectors.

Since we extract many kinds of features, some features might have multi-collinearity and the following analysis including k-means clustering, k-NN, and SVM and NLR might not work well. It is reported that whitening before k-means clustering avoids generating highly correlated cluster centers for obtaining discriminative feature vectors;[30] therefore, we perform whitening. Specifically, to improve the discriminative power of the feature vectors, we make the feature vectors less redundant, i.e., less correlated. To this end, we apply zero-phase component analysis (ZCA) whitening[31] to the feature vectors as follows. Let us denote a matrix, which is obtained by aligning and centering the feature vectors contained in the training dataset, by $\boldsymbol{\Psi} \in \mathbb{R}^{\mathrm{D_t} \times \mathrm{N_t}}$ where D_t and N_t are the dimension and the number of feature vectors, respectively. Note that all data are divided into training and test datasets in the supervised machine learning. The training dataset is used to fit the models, while the test dataset is used for evaluation of the obtained models. The aim of ZCA whitening is to approximate the covariance matrix $\boldsymbol{C} = \boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{T}}$ to the identify matrix $\boldsymbol{I} \in \mathbb{R}^{\mathrm{D_t} \times \mathrm{D_t}}$. To achieve this aim, we first perform the eigenvalue decomposition $\boldsymbol{C} = \boldsymbol{U}\Lambda\boldsymbol{U}^{\mathrm{T}}$, where $\boldsymbol{U}$ and Λ are the eigenvector matrix and diagonal matrix whose diagonal elements are eigenvalues, respectively. Thus, we can obtain the whitening matrix $\boldsymbol{W}_{ZCA} = \boldsymbol{U}\Lambda^{-(1/2)}\boldsymbol{U}^{\mathrm{T}}$. Finally, we can achieve the above aim by transferring the original data $\boldsymbol{\Psi}$ as $\boldsymbol{\Psi}_{ZCA} = \boldsymbol{W}_{ZCA}\boldsymbol{\Psi}$. For the feature vectors contained in the test dataset, we also perform ZCA whitening by using the same matrix $\boldsymbol{W}_{ZCA}$. By transferring the feature vectors via $\boldsymbol{W}_{ZCA}$, the classification accuracy can be improved since the feature vectors become less redundant and have more discriminative power. Note that we standardize the feature vectors after ZCA whitening to obtain the suitable vectors for the subsequent analysis.

D. Multi-stage classification for fish sound detection

A machine learning approach, which integrates results by multiple classifiers to improve the classification performance, works well when classification results are less redundant to each other;[32] therefore, we employ classifiers SVM,[19] *k-NN*,[20] and NLR,[18] which provide classification boundaries that are different from each other. Although each classifier is not necessarily independent, we compared performances to select the appropriate classifier. It should be noted that we classify two classes, i.e., segments that include fish sounds and segments that do not include fish sounds. We define segments that include fish sounds, which were detected by the rule-based filter explained in Sec. II A, as "positive samples." Segments that do not include fish sounds are represented as "negative samples."

First, we prepare a training dataset from the obtained feature vectors. Previous studies[33,34] showed that classifiers cannot be adequately constructed if the training dataset is an imbalanced one in which the number of positive samples and the number of negative samples are greatly different. If positive (negative) samples are larger than negative (positive) samples, we have to select representative positive (negative) samples to make the numbers of positive and negative samples the same. To this end, we apply k-means clustering[35] to all positive (negative) samples and obtain each cluster center. Then we select positive (negative) samples with the shortest Euclidean distances to the obtained cluster centers, which are utilized for the subsequent processing. In this way, the distribution of the whole feature vectors can be approximated by the samples selected via *k*-means clustering.

SVM, *k-NN*, and NLR are described briefly below.

1. SVM (Ref. 19)

SVM finds the optimal hyperplane by optimizing two criteria, i.e., margin maximization and error minimization. By optimizing them in a non-linear high-dimensional feature space, we can construct a classifier that can estimate the probability μ_s of each sample being a positive sample.[36] Note that non-linear classification becomes feasible via the kernel trick while parameters should be tuned for the classification.

2. *k*NN(Ref. 20)

The *k-NN* algorithm performs classification on the basis of majority voting of the *k* closest training samples to each input test sample. In the experiment shown later, we empirically defined *k* as 101. We notice that *k-NN* can provide the probability μ_k of each sample being a positive sample by simply counting the number of samples, which are classified as positive samples among *k* ones. Note that non-linear classification becomes feasible without complicated parameter tuning while classification results tend to be influenced by noisy data.

3. NLR

Even if SVM and *k*-NN cannot realize accurate classification, NLR works well by monitoring the accuracy (reliability) of each classifier and integrating multiple classification results. Note that NLR is based on the previous work[18] that considers multiple feature vectors and unreliable labels. Although the previous work[18] assumes that low-level features are utilized, our NLR adopts output values of SVM and *k-NN* rather than the low-level features, i.e., underwater acoustic features. By using the output values unlike the previous work,[18] we can obtain high-level features for successful classification. The details of NLR are described as follows. In this paper, we call the classifiers "annotators" and denote the annotators SVM and *k-NN* for each channel by $r \in \{s_L, s_R, k_L, k_R\}$, respectively. Based on the paper in Ref. 18, we assign higher weights to annotators that result in accurate classification results and integrate the results by each annotator to obtain the final results. NLR calculates posterior probability as

$$P(j|\boldsymbol{z}, \boldsymbol{W}) = \frac{\exp\left(\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{z}\right)}{\sum_{j'=1}^{J} \exp\left(\boldsymbol{w}_{j'}^{\mathrm{T}}\boldsymbol{z}\right)}, \tag{1}$$

where $\boldsymbol{W} = [\boldsymbol{w}_j]_{j=1}^{J}$ is a matrix that consists of weight vectors $\boldsymbol{w}_j$ and J is the number of classes, i.e., 2. Also, $\boldsymbol{z}$ is a new feature vector obtained via classification results by each annotator. Motivated by the paper in Ref. 37, which showed that $\boldsymbol{z}$ generated from output values rather than elements of feature vectors enables successful classification, we obtain $\boldsymbol{z}$ by aligning output values by each annotator. $\boldsymbol{W}$ is calculated as follows. Let us denote a classification result for the *i*th training sample through an annotator $r \in \{s_L, s_R, k_L, k_R\}$ by $y_i^{(r)}$. Furthermore, we define $s_i^{(r)}$ as 1 if $y_i^{(r)}$ is a correct result and 0 otherwise. In the training phase, for the training dataset $D = [y_i^{s_L}, y_i^{s_R}, y_i^{k_L}, y_i^{k_R}, \boldsymbol{z}_i]_{i=1}^{N}$ (N being the number of training samples), we calculate $\boldsymbol{W}$ by solving the following optimization problem:

$$\boldsymbol{W}_{ML} = \arg\max_{\boldsymbol{W}} \left[\log P(D|\boldsymbol{W}) - \lambda||\boldsymbol{W}||_F^2\right].$$

In this equation, the first and second terms are log likelihood and a regularization term, respectively. Based on the gradient decent method, we can derive the *j*th column of $\boldsymbol{W}_{ML}$, $\boldsymbol{w}_j$, by the following iteration:

$$\boldsymbol{w}_j \leftarrow \boldsymbol{w}_j + \alpha \\ \times \sum_{i\in\{1,2,\ldots,N\}} \sum_{r\in\{s_L,s_R,k_L,k_R\}} \left[s_i^{(r)}\{a_{ij}^{(r)} - P(j|\boldsymbol{z}_i, \boldsymbol{W}\}\boldsymbol{z}_i\right] - 2\lambda\boldsymbol{w}_j.$$

Here, $a_{ij}^{(r)}$ is 1 if $y_i^{(r)} = j$ and 0 otherwise. In the test phase, by using $\boldsymbol{w}_j$ obtained in the training phase and $\boldsymbol{z}$ generated from the test data, we can obtain classification results based on Eq. (1). In this way, NLR is realized on the basis of adaptive feature weighting by focusing on the accuracy of classification results by SVM and *k-NN*.

Since useful classifiers for fish sound detection are different depending on the underwater environment, multiple classifiers should be utilized adaptively. Specifically, we sequentially apply each classifier to audio segments and monitor its classification accuracy, i.e., reliability of the results. If the results are not reliable, the other classifiers are sequentially utilized to accurately detect fish sounds. The details are given below.

a. Classification using SVM. If both probabilities μ_s for two channels are more than a threshold Th_1 or they are less than a threshold Th_2, we regard the results as being reliable; then, the mean of μ_s for two channels is defined as the final result. Otherwise, we do not determine the SVM results as final results and proceed to "b."

b. Classification using k-NN. In the same manner, if both probabilities μ_k for two channels are more than Th_1 or they are less than Th_2, we regard the results as being reliable; then, the mean of μ_k for two channels is defined as the final result. Otherwise, we do not determine the *k-NN* results as final results and proceed to "(c)."

c. Classification using NLR. By using the results of SVM and *k-NN*, NLR is performed to define the probability shown in Eq. (1) as the final result.

TABLE I. Numbers of training and test samples.

Ratio of training samples to all samples	No. of training samples	No. of test samples
1%	423	41962
3%	1271	41114
5%	2119	40266
10%	4238	38147
20%	8477	33908
40%	16954	25431
60%	25431	16954
80%	33908	8477

In the training phase, we determine optimal parameters Th_1 and Th_2 by performing validation using the training data. Concretely, we extract validation data from the training data. Then we change Th_1 and Th_2 from 0 to 1 at intervals of C_{int} and monitor the accuracy, i.e., coincidence ratio of classification results by the proposed hybrid method and ground truths by the rule-based filter shown in Sec. II A. In the test phase, we use Th_1 and Th_2 when the accuracy is maximum to perform the multi-stage classification.

As a consequence of the multi-stage classification, suitable features and classifiers are adaptively selected according to the target audio segments; thus, successful fish sound detection becomes feasible.

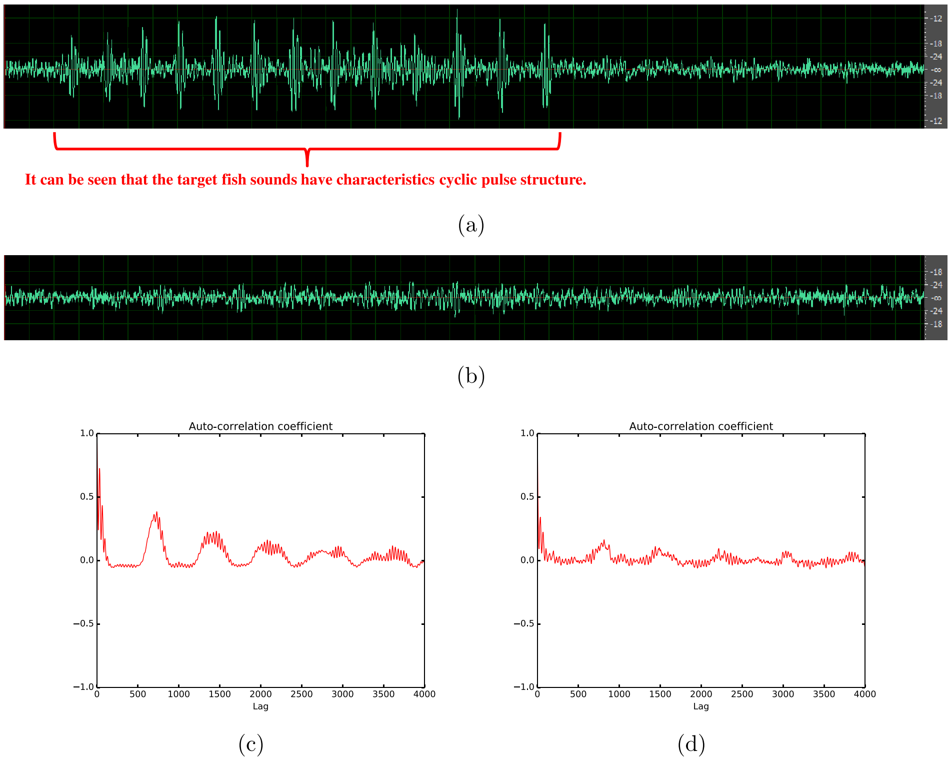
III. RESULTS

A. Parameters used for the analysis

We used underwater acoustic data including sounds of *Sciaenidae* recorded from the coast of Kashima, Japan (see Sec. II A for details). When calculating audio features (see Sec. II C), we set the window width T to 16384 points, i.e., about 0.371 [sec], for adjusting T to the length of the target fish sounds. At the same time, the slide width of the window Δ was set to $T/2$. In this analysis, the parameter for prescreening Th_s (see Sec. II C) was set to 20. According to Sec. II D, we performed our multi-stage classification of two classes, i.e., segments that include fish sounds and segments that do not include fish sounds. From all samples, 1%, 3%, 5%, 10%, 20%, 40%, 60%, and 80% of the samples were randomly selected as training data and the remaining samples were used for test data to verify relations between the number of training samples and detection accuracy. Table I shows the numbers of training and test samples for each case. Note that each sample corresponds to the obtained segments with $T = 16\,384$ points. Furthermore, we randomly divided training data into three sets. The first set was used to train SVM and *k-NN*. For the kernel function in SVM, we used the Gaussian kernel with parameters determined through a grid search.[38] By applying the trained SVM and *k-NN* to the second set, training of NLR was performed. We used the third set as validation data to determine the parameters for our multi-stage classification, Th_1 and Th_2. According to Sec. II D, we performed validation by setting C_{int} to 0.01 and determined Th_1 and Th_2.

B. Evaluations

Based on the above settings, we applied the proposed hybrid method to the constructed dataset. Figure 3 shows examples of results of fish sound detection by the proposed hybrid method. From Figs. 3(c) and 3(d), we can confirm that a fish sound has higher correlation at the constant intervals than an environmental sound; in other words, a fish sound has cyclic pulse structure. We can confirm that the proposed hybrid method can extract the characteristics of fish sounds, i.e., the cyclic pulse structure, which cannot be seen in environmental sounds.



(a)

(b)

(c)

(d)

FIG. 3. (Color online) Examples of fish sound detection results by the proposed hybrid method. The ratio of training samples to all samples was set to 1%. (a) Wave form of an audio segment that was correctly classified as a positive sample. (b) Wave form of an audio segment that was correctly classified as a negative sample. (c) Auto-correlation coefficients for a fish sound. To clarify the pulse structure, auto-correlation coefficients for the square of signal values shown in (a) are calculated. (d) Auto-correlation coefficients for an environmental sound. As in (c), auto-correlation coefficients for the square of signal values shown in (b) are calculated.

Next, we perform quantitative evaluations by comparing among the proposed hybrid method, each single method (SVM, *k-NN*, and NLR) and linear discriminant analysis (LDA).[39] The target species of the proposed hybrid method and the conventional methods[8–14] are not comparable. Since the conventional method[17] is not a supervised machine learning method, we adopted a commonly used classification method, LDA (Ref. 39) for comparisons. LDA maximizes the ratio of between-class variance to the within-class variance. We trained LDA using same training data as SVM and *k-NN* in the proposed hybrid method. In addition, since a shrinkage scheme[40] is useful for improving performance of LDA even for a small number of training samples,[39] we introduced the scheme[40] into LDA. For the evaluation, we define the sensitivity, *a.k.a*, recall and specificity as follows:

$$\text{Sensitivity}\,(\text{Recall}) = \frac{TP}{TP + FN},$$

$$\text{Specificity} = \frac{TN}{TN + FP},$$

where *TP*, *TN*, *FP*, and *FP* are the numbers of true positive, true negative, false positive, and false negative, respectively. Note that *TP* and *TN* are results that are correctly classified as positive and negative samples, respectively, and *FP* and *FN* are results that are mis-classified as positive and negative samples, respectively. Here, we defined test samples including the fish sounds as positive samples, and we defined test samples that do not include the fish sounds as negative samples. Figure 4 shows receiver operating characteristic curves (ROC curves) that display the sensitivity on the vertical axis and the 1-specificity on the horizontal axis, which is a well-known measure in the field of machine learning. We show means of evaluation values when performing random selection of training data five times in this figure. It should be noted that the ROC curve in the top left corner means that the classifier works successfully. It can be seen that classification accuracy differs depending on the classifier. The proposed hybrid method can provide classification accuracy that is equal to or greater than that of other classifiers.

Furthermore, we define the F-measure that evaluates both the comprehensiveness and accuracy of finding positive samples as follows:

$$\text{F–measure} = \frac{2 \times \text{Recall} \times \text{Precesion}}{\text{Recall} + \text{Precision}},$$

where

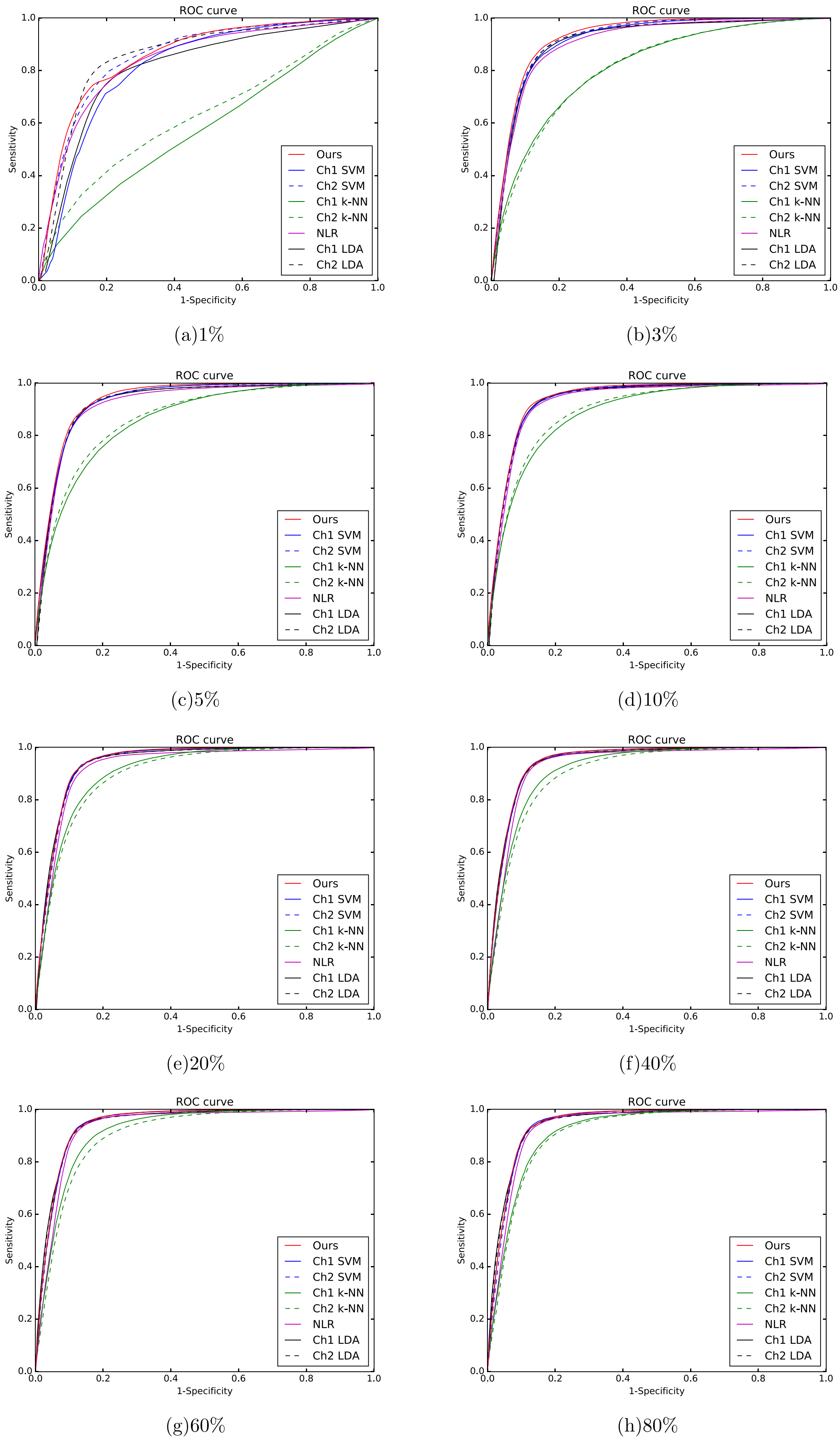
$$\text{Precision} = \frac{\text{No. of correctly classified positive samples}}{\text{No. of positive samples classified by the method}}.$$

Table II shows evaluation values where the F-measure becomes maximum in ROC curves. We show means of evaluation values obtained by repeatedly performing the random selection of training data and evaluation five times. We can see that the proposed hybrid method worked well especially when the ratio of training samples to all samples was small (see 3%, 5%, and 10%). It should be noted that a small number of fish sounds are contained while most of underwater recordings are environmental sounds. Therefore, a large number of positive samples (fish sounds) in training data cannot be prepared unlike other supervised learning tasks such as Web image classification.[41] Even in such a case, the proposed hybrid method can work well by hierarchically obtaining reliable classification results according to the target underwater environment.

IV. DISCUSSION

The merits and limitations of the proposed hybrid method are discussed in this section. First, the proposed hybrid method enables successful fish sound detection even if a single classifier cannot work well (Table II). This merit guarantees that the proposed hybrid method is suitable for a situation in which features and classifiers that provide good detection results differ depending on the underwater environment. Second, we confirmed the merit of the proposed hybrid method compared with existing supervised machine learning, which needs to a large amount of training data. Since there are cases in which underwater acoustic data including many fish sounds cannot be prepared, this merit can contribute to studies such as passive acoustic monitoring in fisheries science. Third, this experiment aims at verifying the effectiveness of the proposed supervised classification shown in Secs. II C and II D by using the croakers' sounds as a case study. For successful classification of the target fish sounds, we performed pre-processing shown in Sec. II C, which are similar to the reference data processing shown in Sec. II A. However, supervised classification shown in Sec. II D can be applied to any fish species.

On the other hand, we can also see the limitation of the proposed hybrid method. Specifically, since the classification boundary may not be suitably learned if the ratio of training samples to all samples is too small (see 1%), the performance was degraded in such a case. Although the proposed hybrid method worked better than comparative methods when the ratio of training samples was small (see 3%, 5%, and 10%), we should investigate the lower limit where the proposed hybrid method can stably provide successful results in the future. Also, there were cases in which the accuracy of a single classifier SVM approached that of the proposed hybrid method as the amount of training data increased. The proposed hybrid method improves the accuracy by compensating for mis-classification by each classifier. Since classification results that are complementary to those by SVM cannot be provided by *k-NN*, this limitation may be caused. In the future, this limitation will be solved by adopting new classifiers that produce classification results that are more independent from each other. Also, we should attempt fish sound detection on other species to verify the generality of the proposed hybrid method. Furthermore, future work should include more detailed analysis based on our fish sound detection method, e.g., estimation of abundance or prediction of relations of fish sounds with seasons, time, and sex.



(a)1%

(b)3%

(c)5%

(d)10%

(e)20%

(f)40%

(g)60%

(h)80%

FIG. 4. (Color online) ROC curves: Means of evaluation values when performing random selection of training data five times are shown. Each caption shows the ratio of training samples to all samples. Results of the proposed hybrid method are denoted by "Ours." "Ch1 SVM," "Ch2 SVM," "Ch1 *k-NN*," "Ch2 *k-NN*," "Ch1 LDA," and "Ch2 LDA" show the results obtained via each classifier, SVM, *k-NN*, and LDA for signals of each channel. "NLR" shows the results obtained by integrating the results of "Ch1 SVM," "Ch2 SVM," "Ch1 *k-NN*," and "Ch2 *k-NN*" via NLR.

TABLE II. Evaluation values when the F-measure becomes maximum in ROC curves. Means of evaluation values when performing random selection of training data five times are shown. Each caption shows the ratio of training samples to all samples. The definitions of "Ours," "Ch1 SVM," "Ch2 SVM," "Ch1 *k-NN*," "Ch2 *k-NN*," "Ch1 LDA," and "Ch2 LDA" are the same as those in the legend of Fig. 4.

(a) 1%

	Ours	Ch1 SVM	Ch2 SVM	Ch1 *k-NN*	Ch2 *k-NN*	NLR	Ch1 LDA	Ch2 LDA
Recall	0.808	0.814	0.810	**0.869**	0.758	0.811	0.793	0.829
Precision	0.707	0.640	0.684	0.429	0.509	0.681	0.655	**0.714**
F-measure	0.753	0.712	0.742	0.563	0.595	0.740	0.717	**0.768**

(b) 3%

	Ours	Ch1 SVM	Ch2 SVM	Ch1 *k-NN*	Ch2 *k-NN*	NLR	Ch1 LDA	Ch2 LDA
Recall	**0.893**	0.878	0.882	0.764	0.764	0.857	0.867	0.880
Precision	0.766	0.754	0.765	0.630	0.615	**0.772**	0.764	0.763
F-measure	**0.824**	0.811	0.819	0.690	0.681	0.812	0.812	0.817

(c) 5%

	Ours	Ch1 SVM	Ch2 SVM	Ch1 *k-NN*	Ch2 *k-NN*	NLR	Ch1 LDA	Ch2 LDA
Recall	**0.904**	0.898	0.903	0.810	0.804	0.884	0.892	0.897
Precision	**0.788**	0.775	0.771	0.663	0.687	0.782	0.781	0.784
F-measure	**0.842**	0.832	0.832	0.729	0.741	0.830	0.833	0.837

(d) 10%

	Ours	Ch1 SVM	Ch2 SVM	Ch1 *k-NN*	Ch2 *k-NN*	NLR	Ch1 LDA	Ch2 LDA
Recall	0.911	0.911	0.910	0.844	0.842	0.902	**0.914**	0.905
Precision	**0.804**	0.794	0.794	0.699	0.712	0.793	0.793	0.797
F-measure	**0.854**	0.848	0.848	0.765	0.771	0.844	0.849	0.847

(e) 20%

	Ours	Ch1 SVM	Ch2 SVM	Ch1 *k-NN*	Ch2 *k-NN*	NLR	Ch1 LDA	Ch2 LDA
Recall	0.922	**0.926**	0.922	0.864	0.853	0.904	0.921	0.919
Precision	**0.803**	0.801	0.797	0.742	0.720	0.796	0.799	0.800
F-measure	0.858	**0.859**	0.855	0.798	0.781	0.847	0.856	0.855

(f) 40%

	Ours	Ch1 SVM	Ch2 SVM	Ch1 *k-NN*	Ch2 *k-NN*	NLR	Ch1 LDA	Ch2 LDA
Recall	0.922	0.923	0.922	0.884	0.862	**0.927**	0.921	0.919
Precision	**0.810**	0.808	0.809	0.752	0.732	0.794	0.807	0.806
F-measure	**0.862**	**0.862**	**0.862**	0.812	0.791	0.856	0.860	0.859

(g) 60%

	Ours	Ch1 SVM	Ch2 SVM	Ch1 *k-NN*	Ch2 *k-NN*	NLR	Ch1 LDA	Ch2 LDA
Recall	0.916	**0.926**	0.923	0.888	0.872	**0.926**	0.925	0.917
Precision	**0.818**	0.811	0.811	0.753	0.729	0.799	0.809	0.811
F-measure	0.864	**0.865**	0.863	0.815	0.795	0.858	0.863	0.861

(h) 80%

	Ours	Ch1 SVM	Ch2 SVM	Ch1 *k-NN*	Ch2 *k-NN*	NLR	Ch1 LDA	Ch2 LDA
Recall	0.922	**0.931**	**0.931**	0.892	0.886	0.924	0.923	0.920
Precision	**0.815**	0.808	0.803	0.746	0.738	0.800	0.810	0.810
F-measure	**0.865**	**0.865**	0.862	0.813	0.805	0.858	0.863	0.862

Finally, we note the comparison of other methods that should be tackled in the future. This paper presents a supervised machine learning method to detect the target fish sounds, which are obtained via the species-specific rule-based filter.[22] To the best of our knowledge, this work is the first attempt to detect the target fish sounds from environmental sounds by a supervised classification scheme. For methods that are applicable to our problem setting, audio classification methods by applying convolutional neural networks (CNN) to the spectrogram have been recently proposed.[42] However, in our problem setting, the number of positive samples in the training data is too few to use CNN and the amount of training and test data is imbalanced. Thus, since fair comparison cannot be conducted even if these methods with are implemented, we compared the proposed hybrid method with sub-techniques that consist of the proposed hybrid method (SVM, *k-NN*, and NLR) and LDA to quantitatively verify the performance of the proposed hybrid method. As a future work, we should develop a new method based on CNN that is applicable to our problem setting and should compare the proposed hybrid method with the developed method.

ACKNOWLEDGMENT

This work was partly supported by JSPS KAKENHI Grant Numbers JP17H01744, JP17H00799, and JST CREST JPMJCR11A1.

[1] P. J. B. Hart and J. D. Reynolds, *Handbook of Fish Biology and Fisheries Volume 2* (Wiley-Blackwell, Hoboken, NJ, 2002).

[2] J. S. Gray, "Marine biodiversity: Patterns, threats and conservation needs," Biodiversity Conserv. **6**, 153–175 (1997).

[3] S. E. Parks, J. L. Miksis-Olds, and S. L. Denes, "Assessing marine ecosystem acoustic diversity across ocean basins," Ecol. Inf. **21**, 81–88 (2014).

[4] L. Hatch, C. Clark, R. Merrick, S. V. Parijs, D. Ponirakis, K. Schwehr, M. Thompson, and D. Wiley, "Characterizing the relative contributions of large vessels to total ocean noise fields: A case study using the Gerry E. Studds Stellwagen Bank National Marine Sanctuary," Environ. Manag. **42**, 735–752 (2008).

[5] M. O. Lammers, R. E. Brainard, W. W. L. Au, T. A. Mooney, and K. B. Wong, "An ecological acoustic recorder (EAR) for long-term monitoring of biological and anthropogenic sounds on coral reefs and other marine habitats," J. Acoust. Soc. Am. **123**, 1720–1728 (2008).

[6] J. L. Miksis-Olds, J. A. Nystuen, and S. E. Parks, "What does ecosystem acoustics reveal about marine mammals in the Bering Sea?," in *The Effects of Noise on Aquatic Life. Advances in Experimental Medicine and Biology*, edited by A. N. Popper and A. Hawkins (Springer, New York, 2012), Vol. 730, pp. 597–600.

[7] S. L. Nieukirk, D. K. Mellinger, S. E. Moore, K. Klinck, R. P. Dziak, and J. Goslin, "Sounds from airguns and fin whales recorded in the mid-Atlantic Ocean, 1999−2009," J. Acoust. Soc. Am. **131**, 1102–1112 (2012).

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

[9] K. Ichikawa, C. Tsutsumi, N. Arai, T. Akamatsu, T. Shinke, T. Hara, and K. Adulyanukosol, "Dugong (Dugong dugon) vocalization patterns recorded by automatic underwater sound monitoring systems," J. Acoust. Soc. Am. **119**, 3726–3733 (2006).

[10] C. Erbe and A. R. King, "Automatic detection of marine mammals using information entropy," J. Acoust. Soc. Am. **124**, 2833–2840 (2008).

[11] E. T. Küsel, D. K. Mellinger, L. Thomas, T. A. Marques, D. Moretti, and J. Ward, "Cetacean population density estimation from single fixed sensors using passive acoustics," J. Acoust. Soc. Am. **129**, 3610–3622 (2011).

[12] T-H. Lin, L-S. Chou, T. Akamatsu, H-C. Chan, and C-F. Chen, "An automatic detection algorithm for extracting the representative frequency of cetacean tonal sounds," J. Acoust. Soc. Am. **134**, 2477–2485 (2013).

[13] D. Diep, H. Nonon, I. Marc, J. Delhom, and F. Roure, "Acoustic counting and monitoring of shad fish populations," in *International AmiBio Workshop: Recent Progress in Computational Bioacoustics for Assessing Biodiversity* (2013), pp. 1–5.

[14] M. Vieira, P. J. Fonseca, M. C. P. Amorim, and C. J. C. Teixeira, "Call recognition and individual identification of fish vocalizations based on automatic speech recognition: An example with the Lusitanian toadfish," J. Acoust. Soc. Am. **138**, 3941–3950 (2015).

[15] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Process. **3**, 72–83 (1995).

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE **77**, 257–286 (1989).

[17] I. Matsuo, T. Imaizumi, and T. Akamatsu, "Detection of fish calls by using the small underwater sound recorder," J. Acoust. Soc. Am. **136**, 2152 (2014).

[18] F. Rodrigues, F. Pereira, and B. Ribeiro, "Learning from multiple annotators: Distinguishing good from random labelers," Pattern Recognit. Lett. **34**, 1428–1436 (2013).

[19] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn. **20**, 273–297 (1995).

[20] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory **13**, 21–27 (1967).

[21] J. Ramcharitar, D. P. Gannon, and A. N. Popper, "Bioacoustics of fishes of the family *Sciaenidae* (croakers and drums)," Trans. Am. Fish. Soc. **135**, 1409–1431 (2006).

[22] T. Lin, Y. Tsao, and T. Akamatsu, "Comparison of passive acoustic soniferous fish monitoring with supervised and unsupervised approaches," J. Acoust. Soc. Am. **143**, EL278–EL284 (2018).

[23] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. (Pearson, London, 2009).

[24] M. Rizwan, B. T. Carroll, D. V. Anderson, W. Daley, S. Harbert, D. F. Britton, and M. W. Jackwood, "Identifying rale sounds in chickens using audio signals for early disease detection in poultry," in *Proc. IEEE Global Conf. Signal and Information Processing* (2016), pp. 55–59.

[25] T. Lim, K. Bae, C. Hwang, and H. Lee, "Classification of underwater transient signals using MFCC feature vector," *Proc. International Symposium on Signal Processing and Its Applications* (2007), pp. 1–4.

[26] P. Du, W. A. Kibbe, and S. M. Lin, "Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching," Bioinformatics **22**, 2059–2065 (2006).

[27] N. Nitanda and M. Haseyama, "Audio-based shot classification for audiovisual indexing using PCA, MGD and Fuzzy algorithm," IEICE Trans. Fundam. **E90-A**, 1542–1548 (2007).

[28] Z. Cataltepe, Y. Yaslan, and A. Sonmez, "Music genre classification using MIDI and audio features," EURASIP J. Adv. Signal Process. **2007**, 36409:1–36409:8 (2007).

[29] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Trans. Speech Audio Process. **5**, 293–302 (2002).

[30] A. Coates and A. Y. Ng, "Learning feature representations with k-means," *Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science* (2012), pp. 561–580.

[31] A. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," Vision Res. **37**, 3327–3338 (1997).

[32] D. Ruta and B. Gabrys, "An overview of classifier fusion methods," Comput. Inf. Syst. **7**, 1–10 (2000).

[33] B. Raskutti and A. Kowalczyk, "Extreme re-balancing for SVMs: A case study," ACM SIGKDD Explor. Newsl. **6**, 60–69 (2004).

[34] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced data sets," Lect. Notes Comput. Sci. **3201**, 39–50 (2004).

[35] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. 5th Berkeley Symposium Mathematical Statistics and Probability* (1967), pp. 281–297.

[36] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," Available at https://www.csie.ntu.edu.tw/~cjlin/libsvm/ (Last viewed October 7, 2018), pp. 1–39.

[37] K. Sasaki, T. Ogawa, S. Takahashi, and M. Haseyama, "DLF-based speech segment detection and its application to audio noise removal for video conferences," ITE Trans. Media Technol. Appl. **4**, 68–77 (2016).

[38] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Technical report, Department of Computer Science (2003).

[39] P. Xu, G. N. Brock, and R. S. Parrish, "Modified linear discriminant analysis approaches for classification of high-dimensional microarray data," Comput. Stat. Data Anal. **53**, 1674–1687 (2009).

[40] O. Ledoit and M. Wolf, "Honey, I shrunk the sample covariance matrix," J. Portfolio Manage. **30**, 110–119 (2004).

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NIPS)* (2012), pp. 1097–1105.

[42] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*(2017), pp. 131–135.