Applied Acoustics 110 (2016) 248–255

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Identification of fish vocalizations from ocean acoustic data

Farook Sattar [a,*], Sarika Cullis-Suzuki [b], Feng Jin [c]

[a] Department of Electrical & Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada
[b] Ocean Networks Canada, University of Victoria, Victoria, British Columbia, Canada
[c] Department of Electrical & Computer Engineering, Ryerson University, Toronto, Ontario, Canada

ARTICLE INFO

Article history:
Received 26 July 2015
Received in revised form 29 December 2015
Accepted 22 March 2016

Keywords:
Ocean acoustics
Fish vocalizations
Identification
High-resolution
Descriptors

ABSTRACT

A new method for identification of fish vocalizations based on auditory analysis and support vector machine (SVM) classification is presented. In this method, high resolution features have been extracted from fish vocalization data using the amplitude modulation spectrogram (AMS) of the input signals to facilitate the identification of grunts and growls made by a highly vocal wild fish, *Porichthys notatus*. The comparison results made from ocean audio recordings verify the effectiveness of the proposed method in identifying various types of fish vocalizations. The relationships between signal-to-noise ratio (SNR) and ocean temperature with the accuracy of the proposed method have also been quantified. Moreover, a context-aware prediction algorithm is introduced for estimating the continuous data.

1. Introduction

Acoustic communication is an important component of intra and inter-specific interactions among many species of fish [1]. Fish produce sound in agonistic situations [2–4], courtship and reproduction events [5–8], and unintentionally during other behaviours [9]. These sounds can range from barely audible to the human ear [10] to loud enough to disturb the sleep of nearby residents [11]. To date, over 800 species are known to make sound and many more are believed to do so [12,13].

Passive acoustics allows a non-destructive way to gain insights on spawning locations, fish abundance, and temporal aspects [12,14]. However, it also relies on the basic recognition of fish sounds, the majority of which to date, have not yet been identified [12]. What's more, once sounds have been identified, sifting through extensive audio datasets can manually become a long and tedious process [15,16]. Manual detections can be too time consuming and error-prone (e.g., due to bias or observer fatigue) to yield accurate results over long datasets [17]. Applying machine learning and automated approaches to long acoustic datasets therefore would further this field markedly.

Here we focus on the plainfin midshipman (*Porichthys notatus*), a highly vocal species of toadfish found along the northeast Pacific [18]. This fish makes four distinct vocalizations: the hum, growl, grunt, and grunt train [5]. Grunts and growls are used in antagonistic encounters with conspecifies, while the hum is produced during reproductive months by alpha males trying to attract females to mate [5,19]. Compared with other species, these fish are fairly well understood, and their call characteristics, well documented [20]. However, an automated approach to quantify and identify their sounds in natural habitats and over long time frames has never been created. Such a tool could offer ecological insights on *P. notatus* populations including abundance, habitat location and range, migratory patterns and call diversity *in situ*.

Traditionally, the identification of animal vocalizations has been done by manually analyzing large recorded datasets [16]. But machine-based algorithms offer a more efficient and potentially effective way to filter through long term acoustic data sets [21–23]. For example, in [15], an identification scheme has been presented for different Orthoptera species by using temporal information such as duration between zero-crossings, shape of the waveform and artificial neural network based multilayer perceptron classifier. Similarly, a complicated method for identification of humpback whales has been introduced in [24] by detecting frequency contour and optimizing multiple parameters. Other identification approaches have been proposed based on frequency-domain information, such as the spectrogram correlation based template matching scheme [16], Kalman filter based contour-tracking scheme [25], contour features based scheme [26], and contour signature based scheme [27]. However, in our study, as the characterizing features of *P. notatus*' grunt and growl signals both fall in the lower frequency range (≈100 Hz), and both sounds are of very short duration, a higher resolution temporal–spectral

* Corresponding author.
E-mail address: farook_sattar@yahoo.com.sg (F. Sattar).

http://dx.doi.org/10.1016/j.apacoust.2016.03.025
0003-682X/

signal representation is therefore desired for accurate identification. In many species of toadfish including *P. notatus*, call frequency is correlated with water temperature [28–31]. Therefore, knowing water temperature can help to predict dominant call frequencies and can thus become a useful parameter in automatic auditory identification schemes.

In this paper, we propose a fish sound identification scheme based on auditory analysis using amplitude modulation spectrogram (AMS). The information containing amplitude modulations of the input signal is analyzed and represented in two-dimensional AMS. The extraction of high-resolution features is performed which is motivated by the results from a neurophysiological experiment on periodicity coding in the auditory cortex [32]. A support vector machine (SVM) classifier is then trained on a large number of pre-selected AMS patterns, and classifies the input signals into grunt and growl classes. It is worth mentioning that the high-resolution features extract the subtle and detailed information and contain more distinctive information than low-resolution features.

2. Method

The proposed identification scheme for fish vocalizations is based on auditory analysis for feature extraction followed by a machine learning algorithm for classification. The overall flowchart of our method is shown in Fig. 1. The hydrophone recordings of fish data are partitioned first into blocks of particular segments. Each 1D data block is then converted into a 2D feature map. A high-resolution feature set (descriptors) is then constructed from the feature maps and used as input to the SVM classifier.

2.1. Data preprocessing

For each of the five days that were analyzed here, five minutes of each hour of a 24-h cycle were processed manually by identifying grunts and growls, thus forming 24 five-minute clips per day. Each five minute spectrogram was then examined manually (visually and audibly) using Audacity 2.0.6 [33] by an expert, who recorded all start and end time stamps (in seconds) for each vocalization. Based on the time stamps of the annotated data, all grunt and growl segments were then extracted, resampled (from 44,100 Hz to 16,000 Hz) and resized into N-sample data blocks ($N = 8192$ here, referring to ≈ 0.5 s) followed by time windowing using N-sample Hamming windows [34]. It should be noted that resampling is usually done to reduce the computational complexity of the method, by running it on a signal sampled at a lower rate. Resizing, meanwhile, is done to save memory by compressing the signal without changing its spectral content [35].

2.2. Feature extraction

We have proposed a high-resolution descriptor (i.e., feature set) for the identification of fish species from their vocalizations. Each input fish data block is first bandpass filtered into 25 subbands by a mel-frequency bank [36]. The envelope of each subband is then obtained by using full-wave rectification followed by decimation with a factor of 3. The decimated envelope signals are subsequently partitioned into segments of 128 samples (0.572 ms) using 50% overlapping, Hamming window. The 256-point fast Fourier transform (FFT) of the zero-padded segments is then calculated. The FFT computes the modulation spectrum in each subband with a frequency resolution of 15.6 Hz. For each subband, the FFT magnitudes are multiplied by 15 uniformly spaced triangular-shaped windows across the 15.6–400 Hz range and summed up to generate 15 modulation spectrum amplitudes representing AMS feature

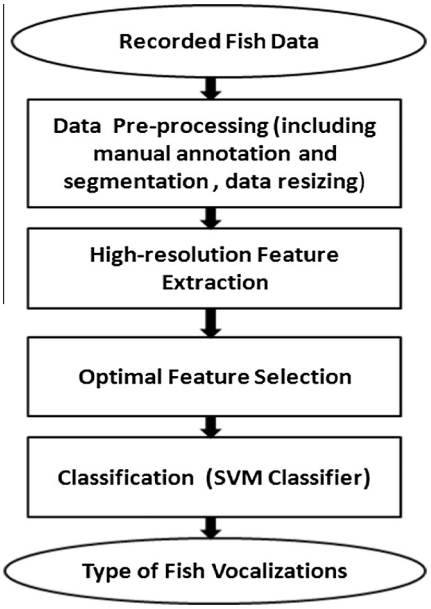

Fig. 1. The overall flowchart of the proposed scheme.

matrix $S_l(n,m)$, where n, m and l indicate the time index, modulation index, and subband/channel index, respectively, with $1 \leqslant \{n,m,l\} \leqslant \{N,M,L\}$. Then, as shown in Fig. 2, the proposed high-resolution descriptor, $\mathbf{d}$ of size $(1 \times ML)$ is constructed as follows:

$$\mathbf{d} = [\Lambda_1(m), \Lambda_2(m), \ldots, \Lambda_L(m)]; \tag{1}$$

where

$$\Lambda_l(m) = \frac{1}{N}\sum_{n=1}^{N} S_l(n,m) \tag{2}$$

Here, we set $M = 25$ and $L = 15$.

Illustrative plots of our high-resolution descriptors for grunt and growl vocalizations are shown in Fig. 3.

2.3. Feature selection

Feature selection is adopted here to improve classification by removing redundant information in high-dimensionality spaces. The sequential floating forward selection (SFFS) algorithm [37,38], finds an optimum subset of features by appending features to and discarding features from subsets of selected features and has been adopted to guide the search, as the SFFS algorithm shows below. A separation index based on distance and separability measures is considered in the SFFS algorithm as an objective function, which evaluates the candidate set by returning a measure of their 'goodness'. This SFFS scheme automatically selects the best feature subset of high-resolution features related to fish vocalizations. The size of the feature space is 375, which corresponds to the length of the high-resolution features.

The SFFS algorithm adopted for feature selection:

1. Start with initialization:
   $i \leftarrow 0$;
   $\mathbf{D}_0 \leftarrow \{\emptyset\}$;
   $J(0) \leftarrow 0$
2. Inclusion – select the most significant feature with respect to $\mathbf{D}_k$:
   $\mathbf{d}' = arg\ \max_{\mathbf{d} \notin \mathbf{D}_k} J(\mathbf{D}_k + \mathbf{d})$;
   $\mathbf{D}_{k+1} = \mathbf{D}_k + \mathbf{d}'; \quad k = k + 1$

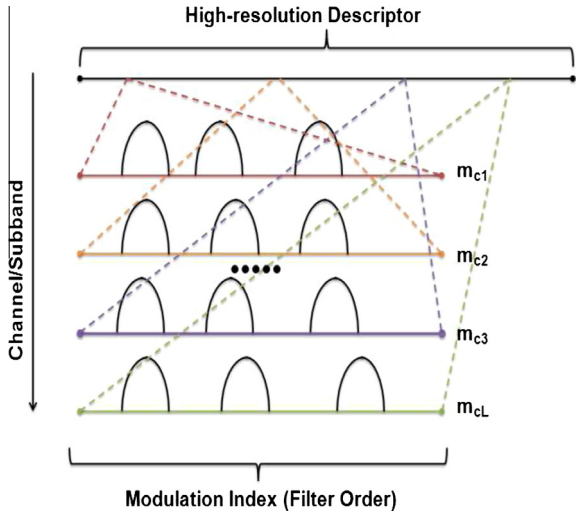

Fig. 2. The construction of the high-resolution descriptor.

3. Conditional exclusion – select the least significant feature in $\mathbf{D}_k$[1]:
   $\mathbf{d}'' = arg\ \max_{\mathbf{d}\in\mathbf{D}_k} J(\mathbf{D}_k - \mathbf{d})$;
4. If $J(\mathbf{D}_k - \mathbf{d}) > J(\mathbf{D}_{k-1})$
   $\mathbf{D}_{k-1} = \mathbf{D}_k - \mathbf{d}''$; $k = k - 1$
   Go to 3
5. Else
   Go to 2
6. End

Techniques for fusing multiple sources of evidence can be generally categorized into two types: fusion at the "feature level" and fusion at the "decision level". Feature-level fusion is performed by merging the calculated features from each source into a cumulative structure and feeding them to a classifier. In decision-level fusion, each feature set is first classified independently, and the final decision is made by fusing the output from the classification processes using the maximum, average, and product criteria. Here, we make use of the feature-level fusion strategy since it often gives better classification accuracy [39].

2.4. Classification

Detecting grunts and growls relating to the specific fish species, is a binary classification problem, which is solved here by a least-squares support vector machine (LS-SVM) [40,41] which has fast convergence, high accuracy, and low computational complexity [42]. LS-SVMs apply linear least squares criteria to the minimization of the cost function instead of traditional quadratic programming. Suppose that the training set consists of W feature vectors $\mathbf{x}_j \in \Re^d (j = 1, 2, \ldots, W)$ from the d-dimensional feature space $\mathbf{X}$, namely, the space representing the features extracted from the principal differential analysis. For each vector $\mathbf{x}_j$, we associate a target $\mathbf{y}_j \in \{-1, +1\}$. The linear SVM classification approach consists of looking for a separation between the two cases in $\mathbf{X}$ by means of an optimal hyperplane that maximizes the separating margin. In the nonlinear case, they are first mapped with a kernel method in a higher dimensional feature space, i.e., $\Phi(\mathbf{X}) \in \Re^{d'} (d' > d)$. The membership decision rule is based on the function sign$[f(x)]$, where $f(x)$ represents the discriminant function associated with the hyperplane in the transformed space and is defined as

[1] Note that book-keeping should be done to avoid infinite loops.

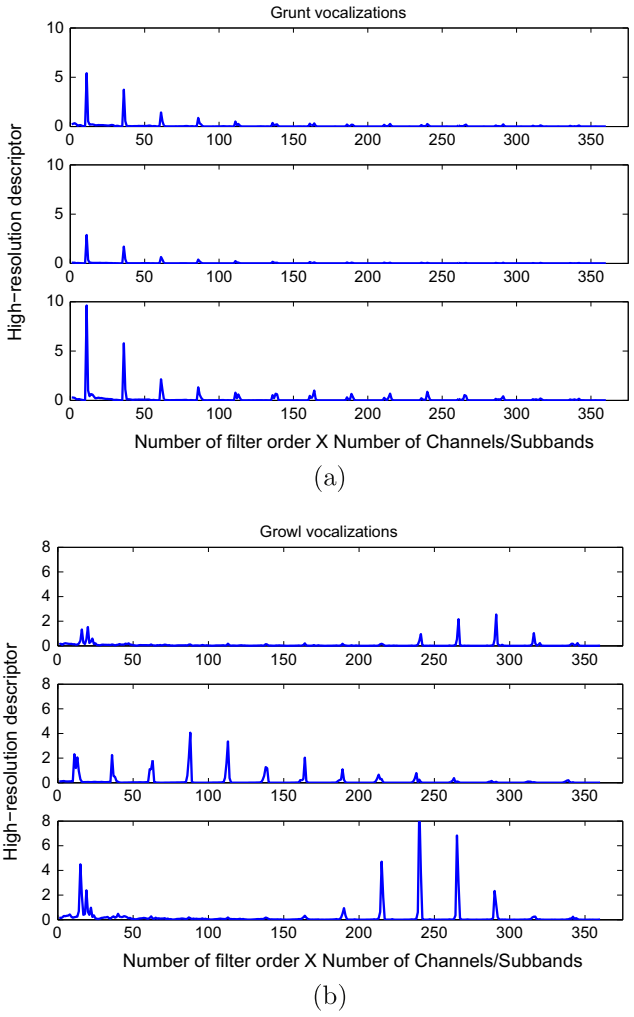

Fig. 3. Examples of high-resolution descriptors for fish vocalizations of (a) grunts and (b) growls, respectively.

$$f(x) = w^* \cdot \Phi(x) = b^* \tag{3}$$

The optimal hyperplane defined by the weight vector $w^* \in \Re^{d'}$ and the bias $b^* \in \Re$ minimizes a cost function that expresses a combination of two criteria: margin maximization and error minimization. It is expressed as

$$\Psi(w, \xi) = \frac{1}{2}||w||^2 + \frac{1}{2}C\sum_{j=1}^{W}\xi_j \tag{4}$$

This cost function minimization is subject to the following constraints:

$$\begin{aligned} &y_j(w \cdot \Phi(x_j) + b) = 1 - \xi_j, \quad j = 1, \ldots, W \\ &\xi_j \geqslant 0, \end{aligned} \tag{5}$$

where ξ_j represents the 'slack' variables introduced to account for nonseparable data. The constant C represents a regularization parameter [43].

3. Experiment

3.1. Experimental dataset

Audio data were collected passively off a private dock located on the east coast of Quadra Island (lat/lon: 50.11159, −125.21757) in June, 2012. Recordings were made with an HTI-96-MIN hydrophone

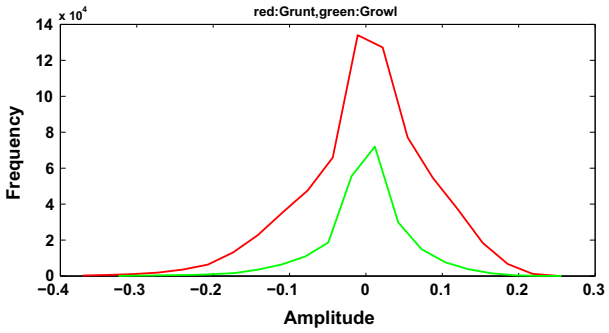

Fig. 4. An illustrative plot of the histograms for grunt/growl data (June 15, 2012).

Table 1
Performance of classification given S proposed features derived from fish data for one hour (e.g., June 15, 2012) (μ: mean, σ: standard deviation).

S	Sensitivity (%) $\mu \pm \sigma$	Specificity (%) $\mu \pm \sigma$	Accuracy (%) $\mu \pm \sigma$
50	100 ± 3.44	90.21 ± 9.68	96.00 ± 3.44
100	100 ± 2.68	91.51 ± 11.32	96.57 ± 3.44
360	100 ± 3.40	90.67 ± 10.19	96.14 ± 3.40

Table 2
Performance of classification given S MFCC and LPCC features derived from fish data for one hour (e.g., June 15, 2012) (μ: mean, σ: standard deviation).

Type of features	S	Sensitivity (%) $\mu \pm \sigma$	Specificity (%) $\mu \pm \sigma$	Accuracy (%) $\mu \pm \sigma$
MFCC	50	73.13 ± 4.44	91.05 ± 1.30	88.28 ± 1.14
	100	73.93 ± 5.81	91.79 ± 1.09	88.36 ± 1.48
	360	74.62 ± 2.52	91.91 ± 0.68	88.86 ± 1.12
LPCC	50	64.17 ± 4.48	92.48 ± 1.32	89.21 ± 1.56
	100	66.79 ± 4.49	92.56 ± 1.14	89.56 ± 1.39
	360	60.86 ± 5.64	89.14 ± 2.08	85.89 ± 1.78

(Wildlife Acoustics, MA, USA) fastened to the bottom of the seafloor, from a depth ranging between 1.5 m and 7.5 m. Five dates were chosen in June for the analysis: June 1st, 7th, 15th, 22nd and 30th. Visual and audio analysis of spectrograms were conducted manually in Audacity 2.0.6 (using Hamming window, 4096-point FFT, and 50% overlap) by taking the first five minutes from each hour of the 24-h cycle, resulting in ten hours of annotated data. Sounds were highlighted and grunt and growl segments were labeled, and start and stop times of vocalizations were recorded, from which total duration was computed. Labels and corresponding durations were then exported as text files. The total number of grunt and growl segments are 4232 and 987, respectively.

3.2. Results and performance

The distributions of the grunts and growls are illustrated in Fig. 4. The completely overlapping histograms of grunt and growl data show the difficulties in grunt/growl identification from time-domain data using merely thresholding.

The proposed scheme is evaluated in terms of classification results for real recorded fish data. Results are obtained over 100 different runs in which the feature sets are split randomly by segment where 2/3 of the data are used for training and 1/3 of the data are retained for testing. In each case, the feature set is normalized to have zero mean and unit standard deviation. The classifier parameters (i.e., the regularization parameter C and the RBF kernel parameter σ) are estimated using cross-validation. Parameters are tuned in two steps. First, a modern global optimization technique, coupled simulated annealing (CSA), determines suitable parameters according to the mean-squared error (MSE) criterion [44]. Second, these parameters are then given to a second optimization procedure (simplex or grid search) to perform a fine-tuning step. Table 1 shows the performance of the proposed scheme in terms of sensitivity, specificity, and accuracy (see Eqs. (6a)–(6c)) given feature set with size S, as derived from the fish data. Here, accuracy (%) increases with S before levelling off.

Table 3
Performances of classification given S features derived from 24-h fish data for various days (μ: mean, σ: standard deviation).

Day	S	Sensitivity (%) $\mu \pm \sigma$	Specificity(%) $\mu \pm \sigma$	Accuracy(%) $\mu \pm \sigma$
June 1, 2012	50	100 ± 1.17	79.58 ± 9.90	94.59 ± 1.98
	100	100 ± 1.03	78.06 ± 8.12	94.68 ± 1.54
	360	100 ± 1.22	77.43 ± 8.85	94.35 ± 1.78
June 7, 2012	50	100 ± 0.47	88.26 ± 8.56	98.54 ± 0.75
	100	100 ± 0.39	85.03 ± 10.05	98.39 ± 0.78
	360	100 ± 0.54	86.59 ± 8.54	98.32 ± 0.78
June 15, 2012	50	100 ± 0.93	67.69 ± 5.39	90.12 ± 1.82
	100	100 ± 1.02	66.98 ± 4.88	89.95 ± 1.65
	360	100 ± 0.88	68.19 ± 5.12	90.19 ± 1.60
June 22, 2012	50	100 ± 0.50	56.92 ± 8.03	94.80 ± 1.10
	100	100 ± 0.46	57.77 ± 7.92	94.85 ± 1.05
	360	100 ± 0.50	55.43 ± 7.29	94.48 ± 0.92
June 30, 2012	50	98.37 ± 1.20	60.12 ± 4.78	83.66 ± 1.88
	100	98.40 ± 1.14	59.58 ± 3.97	83.39 ± 1.61
	360	97.97 ± 1.17	58.03 ± 5.11	83.06 ± 1.82

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6a}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{6b}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{FP}) + (\text{TN} + \text{FN})} \tag{6c}$$

where TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative.

The performances with MFCC (mel frequency cepstral coefficients) [45] features are presented in Table 2. Under the same condition, the sensitivity and accuracies are lower compared with the proposed feature sets depicted in Table 1. Note that the following parameters are used: MFCC window length = 20 ms (320 samples), MFCC window overlapping = 50%. Similarly, the performances with LPCC (linear prediction cepstral coefficients) [46] are presented in Table 2 showing results similar to MFCC feature sets.

In Table 3, the results of sensitivity, specificity, and accuracy are presented using 24-h data for various days in June, 2012. As we can see, the mean accuracies (%) are high above 80% and vary with different days. The sensitivity is very high, while the specificity is relatively low, which was due to the effect of high noise (mainly from boats and *P. notatus* humming) from the original data used.

In Fig. 5, the results of average accuracy (%) are presented over five days in June at a particular time (6 am) when the occurrence of grunts/growls are high as depicted in red,[2] indicating that high performance can be achieved consistently by the proposed method. The results of average accuracy (%) as shown in blue, are presented for five different dates in June based on the 24-h fish data with $S = 50$ when the occurrence of grunts/growls are high over a certain period of time (5 am to 8 pm). Fig. 5 shows that high performance can be achieved by the proposed method for the original noisy

[2] For interpretation of color in Fig. 5, the reader is referred to the web version of this article.

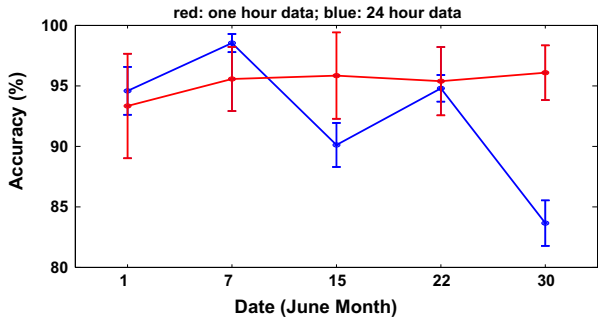

Fig. 5. The error plots of the average classification accuracy (%) over five days in June. Bar corresponds to the standard error from the mean.

Table 4
Performances of classification with feature selection derived from 24-h fish data (μ: mean, σ: standard deviation).

Day	Selected Feature set	Sensitivity (%) $\mu \pm \sigma$	Specificity(%) $\mu \pm \sigma$	Accuracy(%) $\mu \pm \sigma$
June 1, 2012	[7 14 35]	100 ± 1.01	76.89 ± 8.81	94.42 ± 1.56
June 7, 2012	[4 32 35]	100 ± 0.35	88.12 ± 8.16	98.81 ± 0.52
June 15, 2012	[6 11 36]	100 ± 1.24	66.29 ± 5.33	88.30 ± 1.89
June 22, 2012	[7 8 9]	100 ± 0.64	46.36 ± 6.71	93.11 ± 0.99
June 30, 2012	[33 36 43]	97.77 ± 1.15	51.13 ± 3.33	80.20 ± 1.53

data. Note that the number of vocalizations for growls and grunts vary as their ratios are as follows: 84/428 (June 1), 65/854 (June 7), 215/588 (June 15), 158/1344 (June 22), 465/1018 (June 30).

The results of each optimal feature set (using SFFS) are shown in Table 4. As we can see, the performance of the optimal feature sets in 24 h is quite different (e.g., about 18% for June 7 and June 30). This can be explained in terms of the Fisher Discrimination Ratio (FDR) calculated as FDR $= \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}$, where μ_i and σ_i are average sample mean and standard deviation, respectively, of the features for each class label i. For instance, FDR of June 7 data is 0.1526 whereas FDR is 0.0392 for June 30 data, yielding much lower performance (i.e., due to small FDR).

3.3. Statistical analysis

The one-sample, two-tailed t-tests were performed on the results of the optimal feature set (using SFFS) in Table 4 and the corresponding results are presented below:

June 1, 2012 → {accuracy($t(99) = 507.79$, $p < 0.05$, CI = [92.17–92.89]); sensitivity($t(99) = 600.50$, $p < 0.05$, CI = [96.17–96.81]); specificity($t(99) = 80.46$, $p < 0.05$, CI = [71.98–75.62])}

June 7, 2012 → {accuracy($t(99) = 1564.20$, $p < 0.05$, CI = [98.84–99.09]); sensitivity($t(99) = 3127$, $p < 0.05$, CI = [99.60–99.73]); specificity($t(99) = 116.57$, $p < 0.05$, CI = [88.45–91.51])}

June 15, 2012 → {accuracy($t(99) = 513.82$, $p < 0.05$, CI = [87.75–88.43]); sensitivity($t(99) = 763.46$, $p < 0.05$, CI = [96.33–96.83]); specificity($t(99) = 142.37$, $p < 0.05$, CI = [64.86–66.70])}

June 22, 2012 → {accuracy($t(99) = 944.22$, $p < 0.05$, CI = [92.97–93.36]); sensitivity($t(99) = 1707$, $p < 0.05$, CI = [98.67–98.90]); specificity($t(99) = 72.92$, $p < 0.05$, CI = [45.45–48.00])}

June 30, 2012 → {accuracy($t(99) = 474.63$, $p < 0.05$, CI = [78.21–78.86]); sensitivity($t(99) = 665.06$, $p < 0.05$, CI = [93.05–93.60]); specificity($t(99) = 129.08$, $p < 0.05$, CI = [48.47–49.98])}

where CI: confidence interval.

Table 5
SNR(dB) vs call accuracy $\left(SNR_r = \frac{\text{average SNR for Grunt segments}}{\text{average SNR for Growl segments}},\ \mu:\ \text{mean}\right)$.

Day in 2012	SNR(Grunt) μ	SNR(Growl) μ	SNR_r	Accuracy (%) μ
June 1	4.87	1.74	2.79	94.42
June 7	6.63	3.21	2.06	98.81
June 15	10.87	2.77	3.92	88.30
June 22	5.10	1.65	3.09	93.11
June 30	8.15	1.54	5.29	80.20

Table 6
The temperature vs call accuracy with the constraint of specificity (μ: mean).

Day in 2012	Temperature (°C) μ	Accuracy (%) μ	Specificity (%) μ
June 1	13.92	94.42	76.89
June 7	13.81	98.81	88.12
June 15	14.60	88.30	66.29
June 22	16.41	93.11	46.36
June 30	16.49	80.20	51.13

3.4. Relationship between signal-to-noise ratio (SNR) and call accuracy

The average SNR(dB) and the corresponding average call accuracy for five days in June are shown in Table 5. As we can see, the higher the ratio of the average SNR between grunt and growl segments (SNR_r), the lower the call accuracy. The SNRs of each frame are computed for each frequency bin of the STFT (Short term Fourier transform) using window length of 512 samples and FFT points of 512), when the so-called noise tracking algorithm estimates the noise power assuming that the desired signal is "more non-stationary" than the noise [47,48]. Then the SNR of each segment is computed by averaging the SNRs of all frames of a segment.

We found, based on optimal curve fitting, the following relationship between SNR_r and call accuracy, $A(\%)$:

$$A(SNR_r) = \begin{cases} a_0 \cdot SNR_r^0, & SNR_r < 2 \\ a_1 \cdot SNR_r^2 + a_2 \cdot SNR_r + a_3, & SNR_r \in [2, 17] \\ a_4 \cdot SNR_r, & SNR_r > 17 \end{cases} \quad (7)$$

where $a_0 = 100,\ a_1 = -0.06535,\ a_2 = -5.251,\ a_3 = 109.8,\ a_4 = 0.$

3.5. Relationship between temperature, call accuracy, and specificity

The average temperature (°C) and the corresponding average call accuracy for five days in June are shown in Table 6. We found, based on optimal curve fitting, that the relationship between temperature, $T(^\circ\text{C})$, and accuracy, $A(\%)$, follows the *power law* [49] if the specificity, Sp (%) $\geqslant$ 60%, is defined as:

$$A(T) = \begin{cases} 10^{4.04}(T)^{-1.80}, & T \geqslant 13.6\,^\circ\text{C} \\ 10^{2}(T)^{0}, & \text{otherwise} \end{cases} \quad (8)$$

In the case of low specificity (<60%), the above representation of Eq. (8) is scaled by the factor of $\left(\frac{Th}{Sp}\right)$ with $Th = 60$ and the model of accuracy (A) with respect to temperature (T) and specificity (Sp) becomes:

$$A(T, Sp) = \left(\frac{Th}{Sp}\right) \cdot 10^{4.04}(T)^{-1.80}, \quad \forall T \quad (9)$$

3.6. Comparison results

The comparison results are presented in Table 7 based on the method in [15]. Here we can see the classification accuracies are much lower ($\leqslant$70%) than the proposed method (cp. Table 4). Since

Table 7
Comparison results for the 24-h fish data (μ: mean, σ: standard deviation).

Day in 2012	Sensitivity (%) $\mu \pm \sigma$	Specificity(%) $\mu \pm \sigma$	Accuracy(%) $\mu \pm \sigma$
June 1	100 ± 39.16	54.01 ± 49.64	52.58 ± 3.57
June 7	100 ± 25.68	80.94 ± 32.44	61.37 ± 6.66
June 15	100 ± 23.72	86.86 ± 29.93	65.17 ± 8.18
June 22	100 ± 36.81	65.14 ± 43.93	54.22 ± 4.47
June 30	100 ± 34.33	69.81 ± 43.02	57.21 ± 6.26

Table 8
Classification accuracies for the 24-h fish data (μ: mean, σ: standard deviation) with MFCC/LPCC features.

Day in 2012	Accuracy (%) $(\mu \pm \sigma)$ MFCC	Accuracy (%) $(\mu \pm \sigma)$ LPCC
June 1	88.90 ± 2.08	87.66 ± 1.93
June 7	93.11 ± 1.21	93.69 ± 0.86
June 15	86.28 ± 1.30	89.21 ± 1.36
June 22	88.27 ± 0.96	83.12 ± 0.96
June 30	80.06 ± 1.57	80.44 ± 1.44

Table 9
Comparison results for the 24-h fish data (μ: mean, σ: standard deviation) for the HMM based method.

Day in 2012	Sensitivity (%) $\mu \pm \sigma$	Specificity (%) $\mu \pm \sigma$	Accuracy (%) $\mu \pm \sigma$
June 1	45.77 ± 19.65	88.67 ± 18.16	73.63 ± 11.33
June 7	56.08 ± 24.19	90.96 ± 19.31	83.89 ± 7.06
June 15	65.86 ± 25.48	90.57 ± 20.22	71.58 ± 9.64
June 22	70.04 ± 18.70	78.29 ± 25.65	68.03 ± 18.33
June 30	67.22 ± 17.39	61.65 ± 27.13	68.64 ± 15.58

the method in [15] relies on the zero-crossings and the local maximas of the input time-domain signal, it is susceptible to noise and distortion, which gives a lower classification accuracy. Also, the high variabilities in the performance shown in Table 4 indicate lower consistency (i.e., reliability) compared with the proposed method. We have compared the proposed method with that in [15], since so far as we know it provides the best results among the few relevant methods proposed for identification of fish vocalizations.

The classification accuracies for the MFCC/LPCC feature sets + SVM classifier are presented in Table 8. Note that the following parameters are used: MFCC window length = 20 ms (320 samples), number of MFCC features = 12, MFCC window overlapping = 50%, number of LPCC features = 12.

Table 9 shows the performance of the simple and efficient hidden Markov model (HMM) method [50]. The HMM method uses the first 12 MFCC with frames of 20 ms in length and one mixture component per state. Under different conditions (such as temperature, SNR), the proposed method achieves better performance with higher classification accuracies than the HMM based method.

3.7. Cross-validation

The accuracy of the modeled relationship between temperature and call accuracy as depicted in Eqs. (8) and (9), is cross-validated in terms of normalized prediction error (%) between the actual and estimated temperature values. The specificities, sensitivities, and accuracies for all of June 2012 (listed in Table 10) have first been predicted based on the results shown in Table 4. The basic idea is to predict the values for a time interval (e.g., six days), given data of certain intervals (such as every seventh day), based on the concept of collaborative filtering [51] cum super-resolution to reconstruct the mesh grid [52]. The corresponding algorithms which we developed, are presented in Appendix A. Note also that the corresponding value of the control parameter $c = 45$, is chosen empirically, and the parameter, *iter*, referring to the number of iterations, should be equal to the number of days to be predicted between two days. For example, suppose we would like to predict five days between June 1, 2012 and June 7, 2012; we would then set *iter* = 5.

Table 10
Performances of classification predicted and the average temperatures for June, 2012 (μ: mean, σ: standard deviation).

Day	Sensitivity (%) $\mu \pm \sigma$	Specificity (%) $\mu \pm \sigma$	Accuracy (%) $\mu \pm \sigma$	Actual (°C) μ	Predicted (°C) μ	Normalized error (%)
June 1, 2012	100 ± 1.01	76.89 ± 8.81	94.42 ± 1.56	13.92	14.02	0.71
June 2, 2012	100 ± 0.92	78.29 ± 8.72	94.96 ± 1.43	13.95	13.98	0.21
June 3, 2012	100 ± 0.84	79.69 ± 8.64	95.51 ± 1.30	14.18	13.94	1.69
June 4, 2012	100 ± 0.76	81.10 ± 8.56	96.06 ± 1.17	14.23	13.90	2.31
June 5, 2012	100 ± 0.68	82.50 ± 8.48	96.61 ± 1.04	14.10	13.85	1.77
June 6, 2012	100 ± 0.59	83.90 ± 8.40	97.16 ± 0.91	13.92	13.81	0.79
June 7, 2012	100 ± 0.35	88.12 ± 8.16	98.81 ± 0.52	13.81	13.73	0.57
June 8, 2012	100 ± 0.43	86.71 ± 8.24	98.28 ± 0.65	13.57	13.72	1.10
June 9, 2012	100 ± 0.35	88.12 ± 8.16	98.81 ± 0.52	13.74	13.68	0.43
June 10, 2012	100 ± 0.46	85.39 ± 7.80	97.49 ± 0.69	13.29	13.78	3.68
June 11, 2012	100 ± 0.57	82.66 ± 7.45	96.18 ± 0.86	13.62	13.89	1.98
June 12, 2012	100 ± 0.68	79.93 ± 7.09	94.86 ± 1.03	14.10	13.99	0.78
June 13, 2012	100 ± 0.79	77.20 ± 6.74	93.55 ± 1.20	14.24	14.10	0.98
June 14, 2012	100 ± 0.90	74.47 ± 6.39	92.24 ± 1.37	14.38	14.21	1.18
June 15, 2012	100 ± 1.24	66.29 ± 5.33	88.30 ± 1.89	14.60	14.56	0.27
June 16, 2012	100 ± 1.12	69.01 ± 5.68	89.61 ± 1.71	14.60	14.44	1.09
June 17, 2012	100 ± 1.24	66.29 ± 5.33	88.30 ± 1.89	14.60	14.56	0.27
June 18, 2012	100 ± 1.16	63.79 ± 5.50	88.90 ± 1.77	14.71	14.51	1.35
June 19, 2012	100 ± 1.09	61.30 ± 5.67	89.50 ± 1.66	14.82	14.45	2.49
June 20, 2012	100 ± 1.01	58.81 ± 5.84	90.10 ± 1.55	15.66	14.56	7.02
June 21, 2012	100 ± 0.94	56.32 ± 6.02	90.70 ± 1.44	16.19	14.86	8.21
June 22, 2012	100 ± 0.64	46.36 ± 6.71	93.11 ± 0.99	16.41	16.32	0.54
June 23, 2012	100 ± 0.79	51.34 ± 6.36	91.90 ± 1.21	16.65	15.53	6.72
June 24, 2012	100 ± 0.35	48.85 ± 6.53	92.50 ± 1.10	16.67	15.91	4.55
June 25, 2012	100 ± 0.64	46.36 ± 6.71	93.11 ± 0.99	16.57	16.32	1.50
June 26, 2012	100 ± 0.70	46.95 ± 6.28	91.49 ± 1.05	16.73	16.36	1.50
June 27, 2012	99.44 ± 0.76	47.55 ± 5.86	89.88 ± 1.12	17.14	16.41	4.25
June 28, 2012	99.16 ± 0.83	48.14 ± 5.44	88.26 ± 1.19	17.20	16.46	4.30
June 29, 2012	98.88 ± 0.89	48.74 ± 5.02	86.65 ± 1.26	16.56	16.52	0.24
June 30, 2012	97.77 ± 1.15	51.13 ± 3.33	80.20 ± 1.53	16.49	16.79	1.81

Table 11
Classification results for the selected feature sets derived from 24-h fish data (μ: mean,σ: standard deviation) with varying R (ratio of training and testing datasets).

R	Day	Sensitivity (%) $\mu \pm \sigma$	Specificity (%) $\mu \pm \sigma$	Accuracy (%) $\mu \pm \sigma$
1:2	June 1, 2012	98.77 ± 1.11	68.17 ± 9.19	91.86 ± 1.14
	June 7, 2012	100 ± 0.34	87.89 ± 7.51	98.58 ± 0.51
	June 15, 2012	98.94 ± 1.06	62.86 ± 4.23	87.50 ± 1.06
	June 22, 2012	99.70 ± 0.56	45.57 ± 5.42	93.08 ± 0.49
	June 30, 2012	95.87 ± 1.20	47.19 ± 2.98	77.93 ± 0.95
1:3	June 1, 2012	99.15 ± 1.07	67.80 ± 10.46	91.98 ± 1.30
	June 7, 2012	100 ± 0.34	86.34 ± 8.32	98.48 ± 0.59
	June 15, 2012	99.22 ± 1.10	62.02 ± 4.23	87.25 ± 1.00
	June 22, 2012	99.81 ± 0.60	46.22 ± 5.92	93.15 ± 0.45
	June 30, 2012	96.74 ± 1.22	46.11 ± 3.13	77.48 ± 0.98
2:1	June 1, 2012	100 ± 1.01	76.89 ± 8.81	94.42 ± 1.56
	June 7, 2012	100 ± 0.35	88.12 ± 8.16	98.81 ± 0.52
	June 15, 2012	100 ± 1.24	66.29 ± 5.33	88.30 ± 1.89
	June 22, 2012	100 ± 0.64	46.36 ± 6.71	93.11 ± 0.99
	June 30, 2012	97.77 ± 1.15	51.13 ± 3.33	80.20 ± 1.53

The temperature values are then estimated based on the derived relations in Eqs. (8) and (9) using the predicted values in the first three columns of Table 10. Table 10 shows the overall prediction error is as low as 2.14%, which cross-validates the accuracies of the derived relations as well as the prediction scheme.

3.8. Classification results for different ratios of training and testing datasets

The classification results of optimal feature sets (see Table 4) are shown in Table 11 for different ratios of training and testing datasets, R. The results with $R = 1{:}2$ and $R = 1{:}3$ are shown in Table 11 for comparison; note that results corresponding with $R = 2{:}1$ are used throughout this paper. The results appear to be quite consistent for different ratios of training and testing datasets.

4. Conclusion

We have introduced a novel method to identify fish vocalizations using real recorded long term ocean acoustic data. The proposed method, based on auditory analysis (high-resolution descriptor) and SVM classification, demonstrates high classification accuracy in identifying grunts and growls of *P. notatus*. It also outperforms the comparative method using underwater acoustics. Here, the relationship between classification accuracy and SNRs has been shown. Further, the relationship between classification accuracy and water temperature has also been derived and cross-validated based on a context-aware prediction algorithm for estimating the continuous data. Manually segmented data have been adopted for method verification in order to determine the impact of the high resolution features on classification accuracy, independent of possible segmentation errors. A fully automated segmentation method will be developed based on the proposed features in the next step of our work, and investigations into ontology-based annotation and feature enhancement will also be performed. The presented method could be easily adapted for multiclass identification to include different types of vocalizations from other species.

Acknowledgment

The authors are thankful to anonymous reviewers whose critical comments and suggestions helped to improve this paper.

Appendix A. Algorithms

The algorithms corresponding to our context-aware prediction are presented below:

Algorithm for context-aware prediction

Require: $X \in \mathbb{R}^{m \times n}$ {input matrix}
Require: $iter \in \mathbb{R}^{1 \times 1}$ {number of iterations}
1. **while** $j > iter + 1$ not satisfied **do**
2. $Y \leftarrow \mathbf{0}_{(\mathbf{2m-1}) \times \mathbf{n}}$ {zero matrix}
3. $Y(1:2:2m-1, 1:n) \leftarrow X$
4. **while** $i > m - 1$ not satisfied **do**
5. $i \leftarrow i + 1$
6. $Z \leftarrow Y(2i-1:2i+1, 1:n)$
7. Call Code A: $Z' \leftarrow Z$
8. $Y(2i-1:2i+1, 1:n) \leftarrow Z' \in \mathbb{R}^{3 \times n}$
9. **end while**
10. $Y \leftarrow [\mathbf{0}_{2m-1 \times 1}\ Y]$
11. **while** $i > m - 1$ not satisfied **do**
12. $i \leftarrow i + 1$
13. $Z \leftarrow Y(2i-1:2i+1, 1:n)$
14. Call Code A: $Z' \leftarrow Z$
15. $Y(2i-1:2i+1, 1:n) \leftarrow Z' \in \mathbb{R}^{3 \times n}$
16. **end while**
17. $Y \leftarrow [Y\ \mathbf{0}_{4m-1 \times 1}]$
18. **while** $i > m - 1$ not satisfied **do**
19. $i \leftarrow i + 1$
20. $Z \leftarrow Y(2i-1:2i+1, n:n+2)$
21. Call Code A: $Z' \leftarrow Z$
22. $Y(2i-1:2i+1, n:n+2) \leftarrow Z' \in \mathbb{R}^{3 \times n}$
23. **end while**
24. $X \leftarrow Y$
25. **end while**
26. $\tilde{X} \leftarrow X$
Output: $\tilde{X}$ {output matrix}

Algorithm for Code A

Require: $Z \in \mathbb{R}^{3 \times 3}$ {input matrix}
Require: $c \in \mathbb{R}^{1 \times 1}$ {control parameter}
1. $y \leftarrow [Z(1,2)\ Z(2,3)\ Z(3,2)\ Z(2,1)]$
2. $d_1 \leftarrow \mid y(1) - y(3) \mid$
3. $d_2 \leftarrow \mid y(4) - y(2) \mid$
4. $d_3 \leftarrow d_1$
5. $d_4 \leftarrow d_2$
6. $s \leftarrow [d_1\ d_2\ d_3\ d_4]$
7. $b \leftarrow \exp(-s/c)$
8. $y_0 \leftarrow \frac{(b(1)y(1)+b(2)y(2)+b(3)y(3)+b(4)y(4))}{(b(1)+b(2)+b(3)+b(4))}$
9. $Z' \leftarrow Z$
9. $Z'(2,2) \leftarrow y_0$
Output: $Z' \in \mathbb{R}^{3 \times 3}$ {output matrix}

References

[1] Myrberg AAJ, Thresher RE. Interspecific aggression and its relevance to the concept of territoriality in reef fishes. Am Zool 1974;14(1):81–96.
[2] Ladich F. Agonistic behaviour and significance of sounds in vocalizing fish. Mar Freshwater Behav Physiol 1997;29(1–4):87–108.
[3] Myrberg AAJ. Sound production by a coral reef fish (*Pomacentrus partitus*): evidence for a vocal, territorial 'keep-out' signal. Bull Mar Sci 1997;60(3):1017–25.

[4] Bertucci F, Scaion D, Beauchaud M, Attia J, Mathevon N. Ontogenesis of agonistic vocalizations in the cichlid fish *Metriaclima zebra*. CR Biol 2012;335 (8):529–34.
[5] Brantley RK, Bass AH. Alternative male spawning tactics and acoustic signals in the plainfin midshipman fish *Porichthys notatus* Girard (Teleostei, Batrachoididae). Ethology 1994;96:213–32.
[6] Ladich F. Females whisper briefly during sex: context- and sex-specific differences in sounds made by croaking gouramis. Anim Behav 2007;73 (2):379–87.
[7] Lobel PS. Possible species specific courtship sounds by two sympatric cichlid fishes in Lake Malawi, Africa. Environ Biol Fishes 1998;52(443):443–52.
[8] Simes JM, Fonseca PJ, Turner GF, Amorim MCP. African cichlid *Pseudotropheus* spp. males moan to females during foreplay. J Fish Biol 2008;72(10):2689–94.
[9] Kasumyan AO. Sounds and sound production in fishes. J Ichthyol 2008;48 (11):981–1030.
[10] Lobel PS. Sounds produced by spawning fishes. Environ Biol Fishes 1992;33:351–8.
[11] McCosker JE. The Sausalito hum. J Acoust Soc Am 1986;80(6):1853–4.
[12] Rountree RA, Gilmore RG, Goudey CA, Hawkins AD, Luczkovich JJ, Mann DA. Listening to fish: applications of passive acoustics to fisheries science. Fisheries 2006;31(9):433–46.
[13] Fay RR, Popper AN, Webb JF. Fish bioacoustics. In: Webb JF, Popper AN, Fay RR, editors. Introduction to fish bioacoustics. Springer; 2008.
[14] Luczkovich J, Pullinger RC, Johnson S, Sprague M. Identifying sciaenid critical spawning habitats by the use of passive acoustics. Trans Am Fish Soc 2008;137 (2):576–605.
[15] Chesmore ED, Ohya E. Automated identification of field-recorded songs of four British grasshoppers using bioacoustic signal recognition. Bull Entomol Res 2004;94:319–30.
[16] Mellinger DK. A comparison of methods for detecting right whale calls. Can Acoust 2004;32:55–65.
[17] Arthur BJ, Morita TS, Coen P, Murthy M, Stern DL. Multi-channel acoustic recording and automated analysis of *Drosophila* courtship songs. BMC Biol 2013;11.
[18] Arora HL. Observations on the habits and early life history of the Batrachoid fish, *Porichthys notatus* Girard. Copeia 1948;1948(2):89–93.
[19] Bass AH, Bodnar D, Marchaterre M. Complementary explanations for existing phenotypes in an acoustic communication system. In: Hauser M, Marc D, Konishi, editors. The design of animal communication. Cambridge: MIT Press; 1990. p. 493–514.
[20] Bass AH. Sounds from the intertidal zone: vocalizing fish. Bioscience 1990;40 (4):249–58.
[21] Noldus LP, Spink AJ, Tegelenbosch RA. EthoVision: a versatile video tracking system for automation of behavioral experiments. Behav Res Methods Instrum Comput 2001;33(3):398–414.
[22] Delcourt J, Becco C, Vandewalle N, Poncin P. A video multitracking system for quantification of individual behavior in a large fish shoal: advantages and limits. Behav Res Methods 2009;41(1):228–35.
[23] White DJ, Svellingen C, Strachan NJC. Automated measurement of species and length of fish by computer vision. Fish Res 2006;80(2–3):203–10.
[24] Mellinger DK, Martin SW, Morrissey RP, Thomas L, Yosco JJ. A method for detecting whistles, moans, and other frequency contour sounds. J Acoust Soc Am 2011;129(6).
[25] Mallawaarachchi A, Onga SH, Chitre M, Taylor E. Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles. J Acoust Soc Am 2008;124:1159–70.
[26] Esfahanian M, Zhuang H, Erdol N. On contour-based classification of dolphin whistles by type. Appl Acoust 2014;76:274–9.
[27] Ou H, Au WWL, Zurk LM, Lammers MO. Automated extraction and classification of time-frequency contours in humpback vocalizations. J Acoust Soc Am 2013;133:301–10.
[28] Fine ML. Seasonal and geographical variation of the mating call of the Oyster Toadfish *Opsanus tau* L. Oecologia 1978;36:45–57.
[29] Kasumyan AO. Acoustic signaling in fish. J Ichthyol 2009;49(11):963–1020.
[30] Papes S, Ladich F. Effects of temperature on sound production and auditory abilities in the striped raphael catfish *Platydoras armatulus* (Family Doradidae). PLoS One 2011;6(10):26479.
[31] Rubow TK, Bass AH. Reproductive and diurnal rhythms regulate vocal motor plasticity in a teleost fish. J Exp Biol 2009;212:3252–62.
[32] Hall DA, Edmondson-Jones AM, Fridriksson J. Periodicity and frequency coding in human auditory cortex. Eur J Neurosci 2006;24:3601–10.
[33] http://audacity.sourceforge.net/.
[34] Oppenheim AV, Schafer RW. Digital signal processing. Prentice-Hall; 1975.
[35] Proakis J, Manolakis D. Digital signal processing: principles, algorithms and applications. Macmillan Publishing Company; 1992.
[36] Kim G, Lu Y, Hu Y, Loizou PC. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J Acoust Soc Am 2009;126(3):1486–94.
[37] Pudil P, Novovicova J, Kittler JX. Floating search methods in feature selection. Pattern Recogn Lett 1994;15(11):1119–25.
[38] Jin F, Sattar F, Goh DYT. New approaches for spectro-temporal feature extraction with applications to respiratory sound classification. Neurocomputing 2014;123(1):362–71.
[39] Jain A, Nandakumar K, Ross A. Score normalization in multimodal biometric systems. Pattern Recogn 2005;38(12):2270–85.
[40] Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett 1999;9:293–300.
[41] Jin F, Krishnan S, Sattar F. Adventitious sounds identification and extraction using temporal-spectral dominance based features. IEEE Trans Biomed Eng 2011;58(11):3078–87.
[42] Suykens JAK, Vandewalle J, De Moor B. Optimal control by least squares support vector machines. Neural Netw 2001;14(1):23–35.
[43] Joachims T. Making large-scale SVM learning practical. In: Schlkopf B, Burges C, Smola A, editors. Advances in kernel methods-support vector learning. MIT Publisher; 1999.
[44] Salehi SM, Honarvar B. Automatic identification of formation iithology from well log data: a machine learning approach. J Petrol Sci Res (JPSR) 2014;3 (2):73–82.
[45] Devi MR, Ravichandran T. A novel approach for speech feature extraction by cubic-log compression in MFCC. In: IEEE conf on pattern recognition, informatics and mobile eng.
[46] Loizou P. Speech enhancement: theory and practice. New York: Taylor & Francis; 2007.
[47] Ephraim Y, Malah D. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 1984;ASSP-32(6).
[48] Suhadi S, Last C, Fingscheidt T. A data-driven approach to a priori SNR estimation. IEEE Trans Audio Speech Lang Process 2011;19(1).
[49] Bosi M, Goldberg RE. Introduction to digital audio coding and standards. Kluwer Academic Publisher; 2002.
[50] Rabiner LR. A tutorial on hidden markov models and selected applications in speech recognition. Proc IEEE 1989;77:257–86.
[51] Liu H, Hu Z, Mian A, Tian H, Zhu X. A new user similarity model to improve the accuracy of collaborative filtering. Knowl-Based Syst 2014;56:156–66.
[52] Lin C-K, Wu Y-H, Yang J-F, Liu B-D. An iterative enhanced super-resolution system with edge-dominated interpolation and adaptive enhancements. EURASIP J Adv Signal Process 2015;9:1–11.