Applied Acoustics 113 (2016) 193–201

Contents lists available at ScienceDirect

Applied Acoustics

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

Acoustic classification of Australian frogs based on enhanced features and machine learning algorithms

Jie Xie *, Michael Towsey, Jinglan Zhang, Paul Roe

Queensland University of Technology, Australia

ARTICLE INFO

Article history:
Received 23 May 2015
Received in revised form 29 June 2016
Accepted 29 June 2016
Available online 5 July 2016

Keywords:
Frog call classification
Soundscape ecology
Bioacoustics
Acoustic feature extraction

ABSTRACT

Frogs are often considered as excellent indicators of the overall state of the natural environment, but a steady decrease in the frog population has been noticed worldwide. To monitor this change of frog population and optimise the protection policy, frog call classification has become an important bioacoustic research topic. However, automatic acoustic classification of frog calls has not been adequately addressed in the literature. In this paper, an enhanced feature representation for frog call classification using the temporal, perceptual and cepstral features is presented. With the enhanced feature representation, the time-frequency information of frog calls can be effectively represented, which gives a good classification performance. To be specific, each continuous frog recording is first segmented into individual syllables using the Härmä's method. Then, temporal, perceptual, and cepstral features are calculated from each syllable: syllable duration, Shannon entropy, Rényi entropy, zero-crossing rate, averaged energy, oscillation rate, spectral centroid, spectral flatness, spectral roll-off, signal bandwidth, spectral flux, fundamental frequency, linear predictive coding, and Mel-frequency cepstral coefficients. Next, different feature vectors are fused to obtain different enhanced feature representations. Finally, different enhanced feature representations are compared using five machine learning algorithms: linear discriminant analysis, K-nearest neighbour, support vector machines, random forest, and artificial neural network. Experiment results show that our proposed feature representation could achieve better classification performance comparing to other methods with twenty-four frog species, which are geographically well distributed throughout Queensland, Australia.

1. Introduction

Nowadays, great pressure has been placed on global biodiversity due to habitat loss, invasive species, pollution, climate change, and resources overexploitation [1]. Consequently, animal (frog) population has been dramatically decreased. On one hand, frog population is declining, on the other frogs are often regarded as excellent bio-indicators because of their sensitivity to the environment change. Thus, it is becoming ever more necessary to monitor the frog population.

Since frogs are often heard rather than seen[1] and vocalisations of frogs consist of acoustic cues for their communication, acoustic has long been utilised to monitor frog species. There are many types of calls made by frogs, including territorial calls, distress calls, warning calls, release calls, and mating calls [2]. Among them, mating calls are termed as advertisement calls, and can be used to identify frog species. Advertisement calls of species, which are more closely related phylogenetically, are predicted to be more similar than those of distant species [3]. Therefore, acoustic information from advertisement calls can be used for frog call classification.

To monitor frogs' advertisement calls, a traditional field survey method, which requires ecologists to physically visit sites to collect biodiversity data, is both time-consuming and costly. In contrast, recent advances in acoustic sensor techniques provide us a new way to monitor environments over larger spatial temporal scales. But the use of acoustic sensors leads to the rapid growth of acoustic data [4]. Developing semi-automatic or automatic methods for the classification of collected acoustic data by sensors is thus in high demand and attracts a lot of research.

Many studies have investigated the recognition or classification of frog calls. Prior frog call classification system is commonly structured as follows: (1) pre-processing, (2) syllable segmentation, (3) feature extraction, (4) feature fusion, (5) classification. Grigg et al. [5] proposed a system to identify 22 frog species recorded in northern Australia based on peak values (intensity of spectrogram) and

* Corresponding author.
E-mail addresses: j3.xie@student.qut.edu.au (J. Xie), m.towsey@qut.edu.au (M. Towsey), jinglan.zhang@qut.edu.au (J. Zhang), p.roe@qut.edu.au (P. Roe).

[1] https://frogs.org.au/index.html.

http://dx.doi.org/10.1016/j.apacoust.2016.06.029
0003-682X/

Quinlan's machine learning system. Lee et al. [6] introduced a recognition method based on the analysis of spectrogram to classify frog and cricket calls. Mel-frequency cepstral coefficients (MFCCs) of each frame were calculated and averaged as the feature, and linear discriminant analysis (LDA) was used for classifying 30 kinds of frog calls and 19 kinds of cricket calls. Huang et al. [7] extracted spectral centroid, signal bandwidth, and threshold crossing rate as features, and used a K-nearest neighbour (K-NN) classifier and support vector machines (SVM) to classify frog calls. Acevedo et al. [8] used three classifiers, LDA, decision tree (DT), and SVM, for automated classification of bird and amphibian calls. The best average classification accuracy achieved was 94.95%. A method for classifying Australia frogs was proposed by Han et el. [9] where they achieved high accuracy by using hybrid spectral-entropy approach with a K-NN classifier. To utilise the time-varying information, Chen et al. [10] developed a novel feature named multi-stage average spectrum (MSAS) to classify frog calls. Syllable length was first employed for the pre-classification of frog calls; then MSAS was used to perform final classification via template matching. In [11], frog calls were classified using Linear predictive coding (LPC), MFCCs and a K-NN classifier. In [3], Gingras et al. presented a system for the classification of frog genus. This automatic system was built on a SVM model, a K-NN algorithm, and a multivariate Gaussian distribution classifier. Three parameters used were mean values for dominant frequency, coefficient of variation of root-mean square energy, and spectral flux, respectively. Huang et al. [12] developed a method for the classification of anuran vocalisations using fast learning neural-networks. The average classification rate can reach up to 93.4% in average. Bedoya et al. [13] used a fuzzy clustering algorithm (Learning Algorithm for Multivariate Data Analysis) for the recognition of anuran calls. Accuracies between 99.38% and 100% were achieved for two datasets, respectively. However, most features used in the prior work are based on either temporal features, perceptual features, or cepstral features. It is obvious that a combination of three types of features can discriminate a wider variety of species that may share similar characteristics in either temporal, perceptual or cepstral information but not all.

In this study, an enhanced feature representation is proposed for frog call classification, which includes temporal, perceptual, and cepstral features, as an extension of our previous paper [14]. Specifically. After segmenting continuous frog calls into individual syllables. Temporal, perceptual, and cepstral features are extracted from each syllable. Next, different features are fused to obtain the unified feature representation. Finally, the unified feature representation is fed into five machine learning algorithms to perform the task of frog call classification. Twenty-four frog species, which are geographically well distributed throughout Queensland, Australia, are used in this experiment. Experiment results show that our proposed enhanced feature representation can achieve an average classification accuracy of 99.8%, which outperforms other feature representations.

The main contributions and the differences of this work with respect to Xie et al. [14] are (1) the design and realisation of a wide data set of more frog species, with highly noisy background, occurring at different SNRs ranging from −10 dB to 40 dB; (2) a novel feature representation based on feature fusion, which achieves a higher classification accuracy; (3) A post-processing step for syllable segmentation, which reduces the bias introduced by segmentation; (4) five machine learning algorithms are compared to perform the classification; (5) a detailed discussion of various window sizes of MFCCs and perceptual features.

The remainder of this paper is organised as follows: Section 2 describes the methods for frog call classification in detail, which consists of data description, pre-processing, syllable segmentation, feature extraction, feature fusion, and classification. Section 3 reports the experiment results and discussion. The conclusion and future work are offered in Section 4.

2. Architecture of the classification system for frog calls

Our frog call classification system consists of six steps (Fig. 1): data description, syllable segmentation, pre-processing, feature extraction, feature fusion, and classification. Detailed information of each step is shown in following subsections. Different from previous studies [7,14], pre-processing is applied to the segmented syllables rather than continuous recordings.

2.1. Data description

In this study, twenty-four frog species, which are widespread in Queensland, Australia, are selected for experiments (Table 1). All the recordings are obtained from David Stewart's CD with a sample rate of 44.10 kHz and saved in MP3 format [15]. Each recording includes one frog species with the duration ranging from eight to fifty-five seconds.

2.2. Syllable segmentation based on an adaptive end point detection

Each recording is made up of multiple continuous calls of one frog species. For frogs, one syllable is an elementary acoustic unit for classification, which is a continuous frog vocalisation emitted from an individual [7]. In this study, one method built on the Härmä's method is used to perform syllable segmentation for frog calls [16]. The syllable segmentation process is based on the spectrogram, which is generated by applying short-time Fourier transform (STFT) to each recording. For STFT, the window function used is Hamming window with the size and overlap being 512 samples and 25%, respectively. The detail of the segmentation method is described in Fig. 2, which is based on the iterative frequency-amplitude information of the spectrogram. This paper focuses on the evaluation of fused features, but the accuracy of segmentation results can greatly affect the classification performance. To reduce the bias introduced by syllable segmentation, the segmented syllables are further filtered. First, those syllables whose length are smaller than 300 samples are removed. Then, those syllables whose averaged energy are smaller than 15% of the maximum energy and larger than 1.5 times the averaged energy are removed for each frog species [3].

In this study, spectrogram smoothing is optionally applied to the spectrogram before the Härmä's algorithm, because some frog species have large temporal gap within one syllable (see in Fig. 3). As for the smoothing, a Gaussian filter (7×7) is applied to the spectrogram, where the size is set taking into account a trade-off between connecting gaps within one syllable and separating adjacent syllables. The segmentation result after smoothing is shown in Fig. 3. The distribution of syllable numbers after segmentation for all frog species is shown in Fig. 4.

2.3. Pre-processing

Since features play an important role in the classification performance, pre-processing is applied to each syllable to improve the accuracy of feature extraction. The pre-processing of each syllable consists of the following steps:

2.3.1. Pre-emphasise

Some collected frog calls have low amplitude but in the high frequency, which will have an effect on feature extraction of the spectrum at the high frequency end. To enhance those high-frequency components and reduce the low-frequency components,

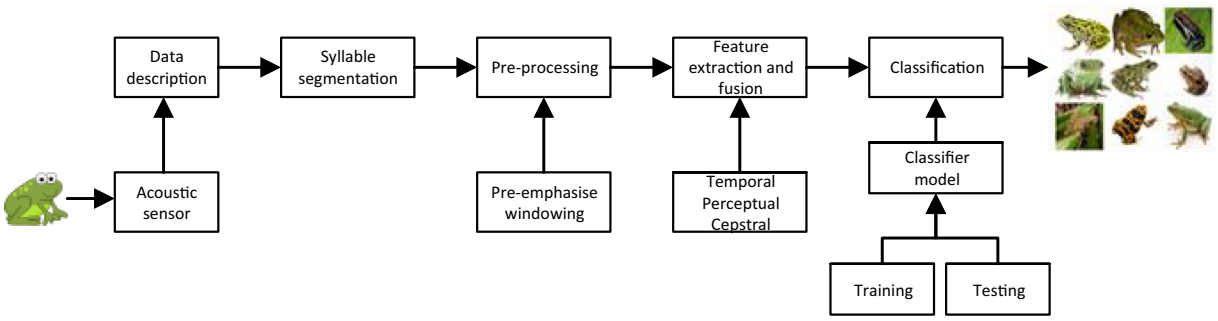


Fig. 1. Flowchart of frog call classification system.

Table 1
Summary of scientific name, common name, and corresponding code. Frog species name with asterisk means that it needs to be smoothed before segmentation.

No.	Scientific-name	Common-name	Code
1	*Assa darlingtoni*	Pouched frog	ADI
2	*Crinia parinsignifera*	Eastern sign-bearing froglet	CPA
3	*Crinia signifera*	Common eastern froglet	CSA
4	*Limnodynastes convexiusculus*	Marbled frog	LCS
5	*Limnodynastes ornatus*	Ornate burrowing frog	LOS
6	*Limnodynastes tasmaniensis**	Spotted grass frog	LTS
7	*Limnodynastes terraereginae*	Northern banjo frog	LTE
8	*Litoria caerulea*	Australian green tree frog	LCA
9	*Litoria chloris*	Red-eyed tree frog	LCS
10	*Litoria latopalmata*	Broad-palmed frog	LLA
11	*Litoria nasuta*	Striped rocket frog	LNA
12	*Litoria revelata*	Revealed tree frog	LEA
13	*Litoria rubella*	Desert tree frog	LRA
14	*Litoria tyleri*	Southern laughing tree frog	LTI
15	*Litoria verreauxii verreauxii*	Whistling tree frog	LVI
16	*Mixophyes fasciolatus*	Great barred frog	MFS
17	*Mixophyes fleayi*	Fleay's barred frog	MFI
18	*Neobatrachus sudelli**	Painted burrowing frog	NSI
19	*Philoria kundagungan*	Mountain frog	PKN
20	*Philoria sphagnicolus**	Sphagnum frog	PSS
21	*Pseudophryne coriacea*	Red-backed toadlet	PCA
22	*Pseudophryne raveni**	Copper-backed brood frog	PRI
23	*Uperoleia fusca**	Dusky toadlet	UFA
24	*Uperoleia laevigata*	Smooth toadlet	ULA

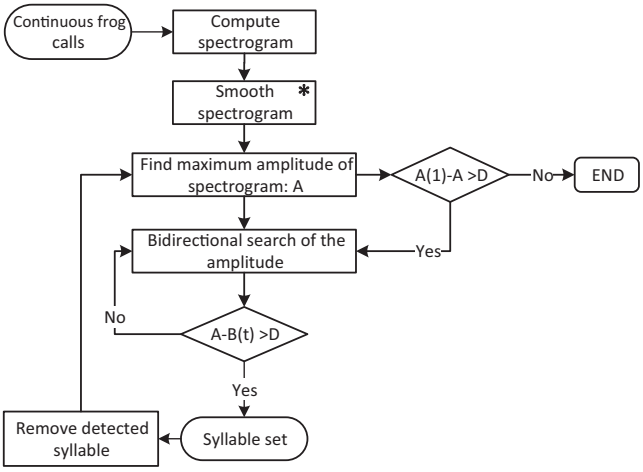
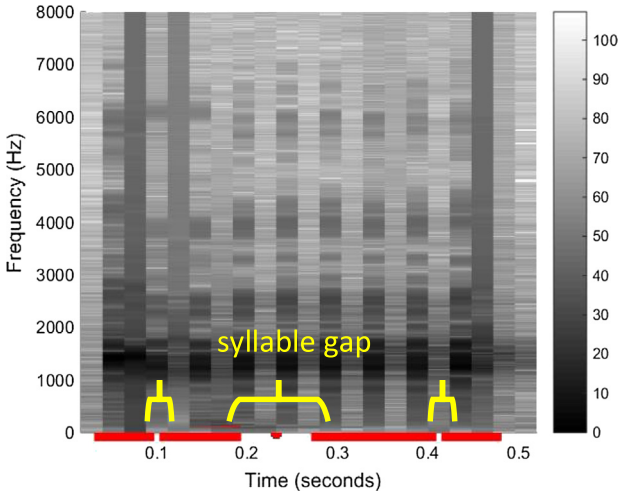
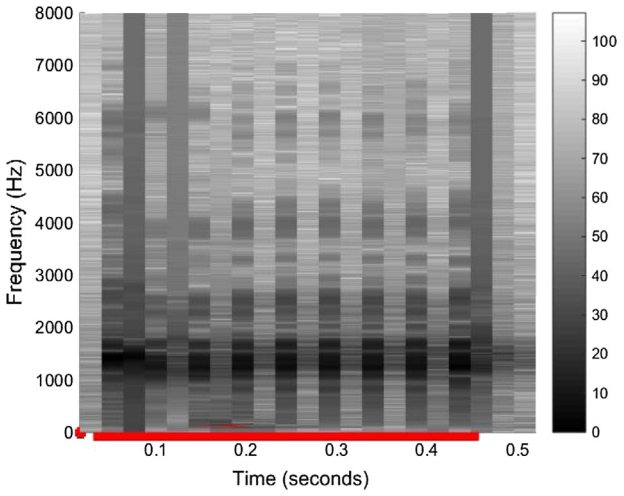


Fig. 2. Segmentation method based on Härmä's algorithm. Here, D is the amplitude threshold for stopping criteria which is set at 20 dB experimentally, and the segmentation result is sensitive with this value. A is the maximum amplitude value of the spectrogram and we save the first maximum amplitude as $A(1)$, $B(t)$ is the amplitude of frame t. An asterisk denotes the optional processing step.



(a) Segmentation result without smoothing



(b) Segmentation result with smoothing

Fig. 3. Syllable segmentation results are marked with a red line for *Neobatrachus sudelli* (one syllable).

a first-order high-pass filter with finite pulse response (FIR) is introduced and defined as follows:

$$y(n) = s(n) - \alpha s(n-1) \tag{1}$$

where $s(n)$ is a syllable of frog call, $y(n)$ is the output of the high-pass filter, α denotes the cut-off frequency of the high-pass filter and is set at 0.97 here, n is the nth sample of the syllable.

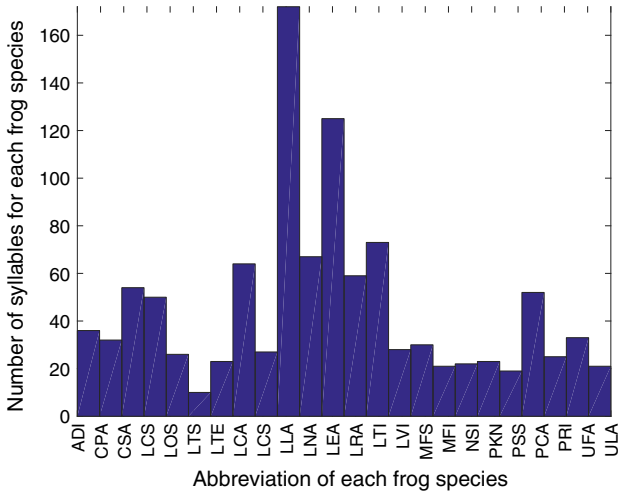


Fig. 4. Distribution of syllable number for all frog species. The *x*-axis is the abbreviation of each frog species, and the corresponding scientific name can be found in Table 1.

2.3.2. Windowing

After pre-emphasising, each syllable is segmented into overlapping frames with fixed length. A Hamming widow is used to minimise the maximum side-lobe in the frequency domain and get side-lobe suppression, which is defined as

$$w(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{L-1}\right), \quad 0 \leqslant n \leqslant L-1 \tag{2}$$

where L is the length of the frame. Because window sizes have an effect on the classification results, different window sizes are optimised for different features in this study. The signal after windowing process is expressed as

$$x(n) = w(n)y(n) \tag{3}$$

2.4. Feature extraction

After pre-processing of each syllable, various parametric representations are used to represent the syllable. In the literature, a variety of parametric representations of frog calls can be found, such as LPC and MFCCs [11,17,13]. MFCCs often achieves a better classification performance than LPC [11]. Different from the hybrid features used in [7,9,3,14], our enhanced feature consists of more features, such as oscillation rate [14], to further improve the classification accuracy. In this study, temporal features include syllable duration, Shannon entropy, rényi entropy, zero-crossing rate, averaged energy, and oscillation rate. Perceptual features contains spectral centroid, spectral flatness, spectral roll-off, signal bandwidth, spectral flux, and fundamental frequency. The MFCCs feature is used as a cepstral feature. The description of each feature is listed below:

(1) syllable duration (*Dr*): Syllable duration [14] is directly obtained from the bounds (time domain) of the segmentation results.

$$Dr = x(n_e) - x(n_s) \tag{4}$$

where n_e and n_s are the end and start location of one segmented syllable.

(2) Shannon entropy (*Se*): Shannon entropy is the expected information content of a sequence of a signal. It is often used to describe the average of all the information contents weighted by their probabilities p_i.

$$Se = -\sum_{i=1}^{L} p_i \log_2(p_i) \tag{5}$$

where L is the length of a frog syllable.

(3) Rényi entropy (*Re*): rényi entropy is calculated to obtain the different averaging of probabilities via the parameter α, and defined as

$$Re = \frac{1}{1-\alpha}\log_2\left(\sum_i^n p_i^{\alpha}\right) \tag{6}$$

where p_i is the probabilities of the occurrence x(n) in the signal.

(4) Zero-crossing rate (*Zcr*): zero-crossing rate denotes the rate of signal change along a signal. When adjacent signals have different signs, a zero-crossing occurs. The mathematical expression of *ZCR* can be defined as

$$Zcr = \frac{1}{2}\sum_{n=0}^{L-1}[sgn(x(n)) - sgn(x(n+1))] \tag{7}$$

(5) Averaged energy (Ae): averaged energy is defined as the sum of intensity of signal.

$$Ae = \frac{1}{L}\sum_{n=0}^{L-1} x(n)^2 \tag{8}$$

(6) Oscillation rate (*Or*): oscillation rate is calculated in the frequency boundary around the fundamental frequency. First, the power within the frequency boundary is calculated. After normalising the power, the first and last 20% part of the power vector is discarded due to the uncertainty. Next, the autocorrelation is performed by the length of the vector. Furthermore, a discrete cosine transform is employed to the vector after mean subtraction, and the position of the highest frequency is achieved to calculate the oscillation rate. Detailed description can be found in our previous study [14].

(7) Spectral centroid (*Sc*): spectral centroid is the centre point of spectrum distribution. In terms of human audio perception, it is often associated with the brightness of the sound. With the magnitudes as the weight, it is calculated as the weighted mean of the frequencies.

$$Sc = \frac{\sum_{k=0}^{N-1} f_k X(k)}{\sum_{k=0}^{N-1} X(k)} \tag{9}$$

where $X(k)$ is the discrete Fourier transform (DFT) of the syllable signal of the kth frame, N is the half size of DFT.

(8) Spectral flatness (*Sf*): spectral flatness provides a way to quantify the tonality of a sound. A higher spectral flatness indicates a similar amount of power of the spectrum in all spectral bands. Spectral flatness is measured by the ratio between the geometric mean and the arithmetic mean of the power spectrum and defined as

$$Sf = \frac{\sqrt{\frac{1}{N}\sum_{k=0}^{N-1} InX(k)}}{\frac{1}{N}\sum_{k=0}^{N-1} X(k)} \tag{10}$$

(9) Spectral roll-off (*Sr*): spectral roll-off is often used to measure the spectral shape, and defined as the frequency H. Here H is the value below which θ of the magnitude distribution is concentrated.

$$\sum_{k}^{H} X(k) = \theta \sum_{k=1}^{N-1} X(k) \tag{11}$$

where θ is set at 0.85.

(10) Signal bandwidth (*Bw*): signal bandwidth can be used to represent the difference between the upper and lower cut-off frequencies.

$$Bw = \sqrt{\frac{\sum_{k=0}^{N-1}(k-Sc)^2|x(n)|}{\sum_{k=0}^{N-1}X(k)}} \tag{12}$$

(11) Spectral flux (*Sf*): spectral flux is used to measure how quickly the power spectrum of a signal is changing. The spectral flux can be obtained via the power spectrum comparison between one frame and its previous one. The calculation of spectral flux is denotes as

$$Sf = \sum_{k=-\frac{N}{2}}^{k=\frac{N}{2}} H[|X(n,k)| - |X(n-1,k)|] \tag{13}$$

where $H(x) = (x + |x|)/2$ is a half-wave rectifier function.

(12) Fundamental frequency: fundamental frequency is calculated via averaging peak intensity of all frames within one frog syllable. If the peak intensity value is higher than an empirically chose or specified threshold, the frequency of that peak will be selected to calculate the fundamental frequency.

(13) Linear prediction coding (LPC): LPC is often used to represent the spectral envelope of speech sound [18]. LPC coefficients can be calculated using a linear predictive filter.

$$X(n) = \sum_{i}^{p} a_i x(n-i) \tag{14}$$

where p is the order of the polynomial a_i. In the proposed study, 13 LPC coefficients are calculated. The value of p is 12 (12th-order polynomial).

(14) Mel-frequency cepstral coefficients (MFCCs): MFCCs, which are obtained by applying discrete cosine transform to a sub-band Mel-frequency spectrum within a short time, have been widely used in bird classification [19], speech/speaker recognition [20], and frog identification [13]. In this study, MFCCs are calculated based on the method of [19].

Step 1: Band-pass filtering: The amplitude spectrum is then filtered using a set of triangular band-pass filters.

$$E_j = \sum_{k=0}^{N/2-1} \phi_j(k) A_k, \quad 0 \leqslant j \leqslant J-1 \tag{15}$$

where J is the number of filters, ϕ_j is the jth filter, and A_k is the amplitude of $X(k)$.

$$A_k = |X[k]|^2, \quad 0 \leqslant k \leqslant N/2 \tag{16}$$

Step 2: Discrete cosine transform: MFCCs for the ith frame are computed by performing DCT on the logarithm of E_j.

$$C_m^j = \sum_{j=0}^{J-1} \cos\left(m\frac{\pi}{J}(j+0.5)\right)\log_{10}(E_j), \quad 0 \leqslant m \leqslant L-1 \tag{17}$$

where L is the number of MFCCs.

In this study, the filter bank consists of 40 triangular filters, that is $J = 40$. The length of MFCCs of each frame is 12 ($L = 12$). After calculating MFCCs of each frame, the averaged MFCCs of all frames within one syllable are calculated.

$$f_m = \frac{\sum_{i=1}^{K}(C_m^l)}{K}, \quad 0 \leqslant m \leqslant L-1 \tag{18}$$

where f_m is the mth MFCCs, K is the number of frames within the syllable.

For all perceptual features and *Zcr*, the mean values are calculated to characterise the frog syllable. Then, the L-dimensional MFCC vectors are fused with the other 11 feature vectors to form the enhanced temporal, perceptual and cepstral (*TemPerCep*) features.

After the formulation of feature vectors, the normalisation is conducted as follows

$$v_i = \frac{v_i - \mu_i}{\sigma_i} \tag{19}$$

where μ_i and σ_i are the mean and standard deviation computed for each feature vector i.

Let F_1 represent temporal features with length L_1, F_2 and F_3 represent perceptual features and cepstral features with length L_2 and L_3, respectively. The enhanced procedure is performed as

$$F_H = w_1 F_1 \oplus w_2 F_2 \oplus w_3 F_3 \tag{20}$$

where w_1, w_2, and w_3 are the weights, $\oplus$ is the concatenation operation.

2.5. Classifier description

In this paper, the classification results for five machine learning algorithms are reported: (1) Linear discriminant analysis (LDA), (2) K-nearest neighbour, (3) Support vector machines, (4) Random forest, (5) Artificial neural network. Five feature vectors, linear predictive coding, MFCCs, enhanced temporal feature and MFCC (*TemCep*), enhanced temporal and perceptual features (*TemPer*), enhanced temporal, perceptual features, and MFCC (*TemCepPer*), are fed into each machine learning algorithm respectively to test their classification performance.

2.5.1. Linear discriminant analysis

After transforming the feature vector into low-dimensional space, the classification accuracy can be improved for linear discriminant analysis (LDA). In LDA, the goal is to find an optimal transformation matrix to transform the feature vector from an n-dimensional space to a d-dimensional space. A linear mapping, which maximises the Fisher criterion J_F, is used to obtain the transformation matrix as follow.

$$J_F(A) = tr((A^T S_w A)^{-1}(A^T S_B A)) \tag{21}$$

where S_w and S_B are the within-class scatter matrix and between-class scatter matrix, respectively. The within-class scatter matrix and between-class scatter matrix are respectively defined as

$$S_W = \sum_{j=1}^{C}\sum_{i=1}^{N_j}\left(F_i^j - \mu_j\right)\left(F_i^j - \mu_j\right)^T \tag{22}$$

$$S_B = \sum_{j=1}^{C}(\mu_i - \mu)(\mu_i - \mu)^T \tag{23}$$

where F_i^j is the ith feature vector of frog species j, μ_j is the mean vector of species j, C is the number of frog species, and N_j is the number of feature vectors in species j, μ is the mean vector of all frog species.

The optimisation of the transform matrix can be determined via finding the eigenvectors of $S_W^{-1}S_B$.

$$A_{opt} = \text{argmax}\frac{tr(A^T S_B A)}{A^T S_W A} \tag{24}$$

In the recognition stage, the feature vector is first transformed into a lower-dimensional space via A_{opt} derived by LDA. Then, the distance between the feature vector of the test syllable and the feature vector representing this species is calculated. The one with minimum distance is regarded as the identified species.

2.5.2. K-nearest neighbour

For the K-NN classifier, the distance between an input frog feature vector and all stored feature vectors is first calculated. Then K closest vectors are selected to determine the species of the input feature vector by majority voting. For example, the Euclidean distance between an input instance i (frog feature vector) and one stored instance j is calculated as

$$d(i,j) = \sqrt{\sum_{c=1}^{n}(F_{i,c} - F_{j,c})^2} \tag{25}$$

Then the species of this input instance i can be predicted from the selected k nearest neighbours. If

$$\frac{1}{k_1}\sum_{j\epsilon S_1} d(i,j(S_1)) \leqslant \frac{2}{k_2}\sum_{j\epsilon S_2} d(i,j(S_2)) \tag{26}$$

where $k = k_1 + k_2, k_1$ is the number of frog species S_1, k_2 is the number of frog species S_2. Here the input instance i will be classified as frog species S_2. Following prior work [9,14], the distance function used for K-NN is the Euclidean function, and k is set at 1.

2.5.3. Support vector machines

Due to the high accuracy and superior generalisation properties, support vector machines (SVM) have been widely used for classifying animal sounds [7,8]. In this study, the feature set obtained is first selected as training data. Then, the pairs $(F_l^n, L_l^n), l = 1, 2, \ldots, C_l$ are constructed using the selected training data, where C_l is the number of frog instance in the training data, F_l^n is the feature vector obtained from the lth frog instance in the training data, L_l^n is the frog species label. Furthermore, the decision function for the classification problem based on SVM [21] is defined by the training data as follows.

$$f(v) = sgn\left(\sum_{sv}\alpha_l^n L_l^n K(v, v_l^n) + b_l^n\right) \tag{27}$$

where $K(.,.)$ is the kernel function, α_l^n is the Lagrange multiplier, and b_l^n is the constant value. In this work, the Gaussian kernel is selected as the kernel function. Parameters α and v are selected independently for each feature vector by grid search using cross-validation [22].

2.5.4. Random forest

Random forest (RF) is a tree-based algorithm, which builds a specified number of classification trees without pruning. The nodes are split on a random drawing of m features from the entire feature set M. A bootstrapped random sample from the training set is used to build each tree. The advantage of RF is its ability to generate a metric to rank predictors based on their relative contribution to the model's predictive accuracy [23]. The prediction is defined as follows.

$$Pred = \frac{1}{K}\sum_{n=1}^{K} T_i \tag{28}$$

where T_i is the nth tree response of the RF. In this work, the number of trees K is set at 300 trees to characterise frog calls. As for the predictor variables m, it is set at $\sqrt{N}$, where N is the feature dimension in a syllable.

2.5.5. Artificial neural network

Artificial neural network (ANN) is a non-linear, adaptive, machine learning tool with great capabilities for learning, generalisation, non-linear approximation, and classification. An ANN architecture often consists of many interconnected neurons organised in successive layers: pattern layer, summation layer, and decision layer. The neuron in class is often computed by a Gaussian function. Then, the summation layer uses summation units to memorise the class conditional probability density functions of each class through a combination of Gaussian densities. Lastly, the decision layer unit classifies the pattern in accordance with the Bayesian decision rule based on the output of all summation layer neurons as follows.

$$D(F) = \text{argmax}p_i(F), i = 1, \ldots, N \tag{29}$$

where i is the species index, N is the total number of frog species.

$$p_i(F) = \sum_{j=1}^{m_i}\beta_{ij}\phi_{ij}(F) \tag{30}$$

where m_i is the number of Gaussian components, β_{ij} and $\phi_{ij}(F)$ can be represented as follows.

$$\sum_{j=1}^{m_i}\beta_{ij} = 1 \tag{31}$$

$$\phi_{ij}(F) = \frac{1}{(2\pi)(d/2)\sigma^d}\exp\left[-\frac{(F-\mu_{ij})^T(F-\mu_{ij})}{2\sigma^2}\right] \tag{32}$$

where $i = 1, \ldots, N, j = 1, \ldots, m_i, d$ denotes the dimension of the input vector F, σ is the smoothing parameter, μ_{ij} is the mean vector and the central of the classification. In this study, one ANN classifier named multiple perception layer (MLP) is used to classify frog calls.

3. Experiment results

In this experiment, performance statistics are estimated with fivefold cross validation. The performance of the proposed frog call classification system is evaluated by quantitatively expressed detection metrics, such as average accuracy, precision, and specificity. The definition of accuracy, precision, and specificity can be defined as

$$Sensitivity = \frac{TP}{TP + FN} \tag{33}$$

$$Specificity = \frac{TN}{TN + FP} \tag{34}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{35}$$

where TP is true positive, FP is true positive, TN is true negative, and FN is false negative.

3.1. Effects of different machine learning algorithms

Fig. 5 shows the frog call classification performance with different machine learning algorithms. The high classification results in term of the accuracy, sensitivity and specificity measure of different machine learning algorithms indicate good classification performance. It can be observed that RF achieves the best classification performance, while the classification performance of LDA is the lowest. Meanwhile, the classification performances

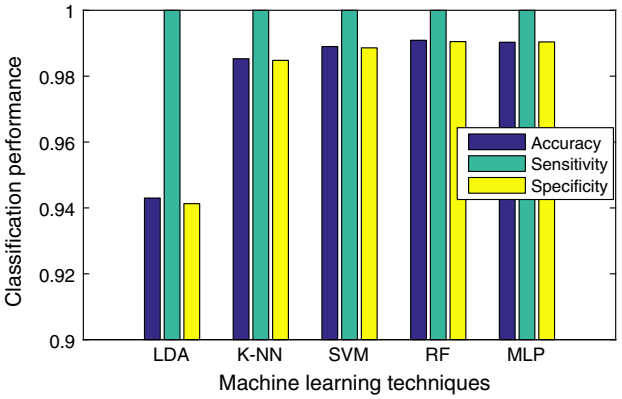


Fig. 5. Classification results of different machine learning algorithms.

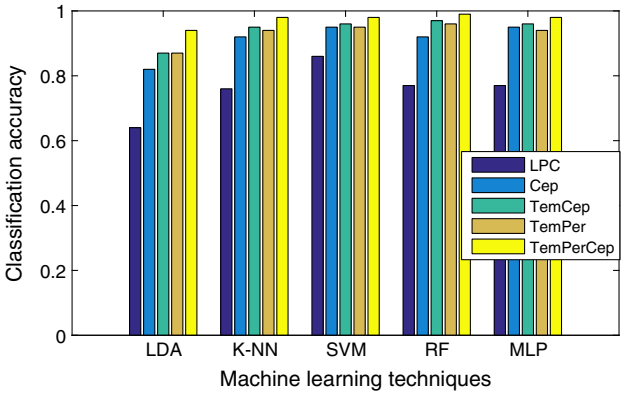


Fig. 6. Classification results with different feature representations.

of SVM and MLP are very good, which might be that the features and machine learning algorithms are quite suitable. It can be seen from Fig. 5 that frog call classification with different machine learning algorithms can achieve good performance with our enhanced feature representation, because the classification accuracy is very high. It can also be noted that RF can be highly recommended for classification of frog calls due to the highest classification accuracy.

3.1.1. Effects of different feature representations

Fig. 6 illustrates the classification accuracy with different feature representations: LPC, MFCCs (*Cep*), temporal features and MFCCs (*TempCep*), temporal features and perceptual features (*TemPer*), and temporal features, perceptual features and MFCCs (*TemPerCep*). It can be seen that cepstral features (*Cep*, *TempCep*, *TemPerCep*) have a more stable performance than LPC and perceptual features. It is evident that our proposed enhanced feature representation (*TemPerCep*) shows outstanding performance of all proposed feature representations of all machine learning algorithms. The reason for the high classification accuracy is that frog calls are of short duration and cover a small spectral band. Our proposed enhanced feature, *TemPerCep*, can better characterise the content of frog calls. Although the classification performance of *TemPerCep* is not significantly higher than other feature representations, the difference does show that our proposed feature representation is suitable and effective for the classification of frog calls.

3.1.2. Effects of different window sizes for MFCCs and perceptual features

Since the window size has an effect on the MFCCs and perceptual features, different window sizes will lead to a different classification performance (Figs. 7 and 8). The window sizes used for test are the 32 samples, 64 samples, 128 samples, 256 samples, respectively, because the syllable length of some frog species is less than 512 samples. It is found that the best classification performance for MFCCs is achieved with window size of 64 samples. For *TemPer*, the window size of 64 samples obtains the best classification performance. It also can be observed that SVM and RF achieve the best classification performance for MFCCs and *TemPer*. Moreover, different window sizes of MFCCs have a larger variation than *TemPer* features, which might be because temporal features have a high weight in *TemPer* for the classification task.

3.1.3. Effects of background noise

To further evaluate the robustness of our proposed feature representation, white noise with different signal-to-noise (SNR) of 40 dB, 30 dB, 20 dB, 10 dB, 0 dB, and −10 dB is added to the frog calls. Because this paper focuses on the evaluation of features

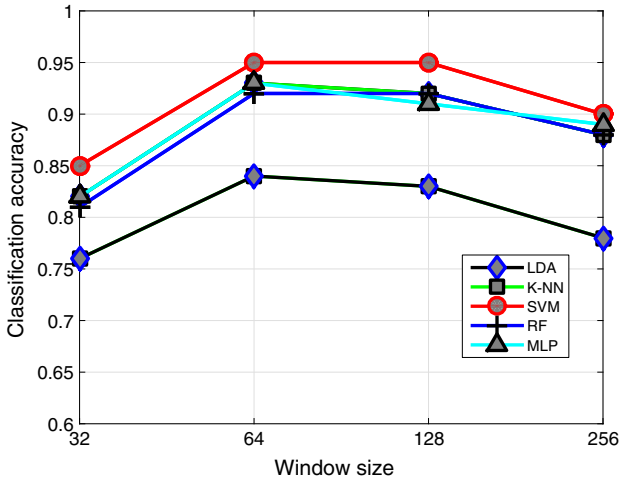


Fig. 7. Classification results of MFCCs with different window sizes.

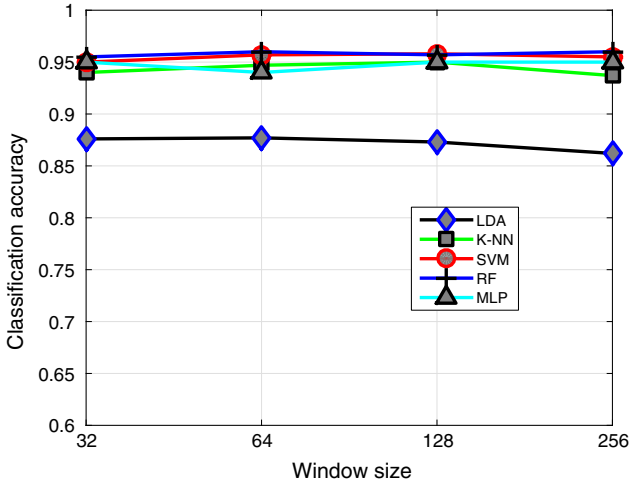


Fig. 8. Classification results of *TemPer* with different window sizes.

rather than the segmentation method, the artificial noise is added after syllable segmentation. Since SVM has shown a good performance and been widely used for frog call classification [8,7], we only use SVM to test the effects of different levels of artificial noise. The classification results of different levels of noise contamination are shown in Fig. 9. It is found from Fig. 9 that MFCCs (Cep) are very

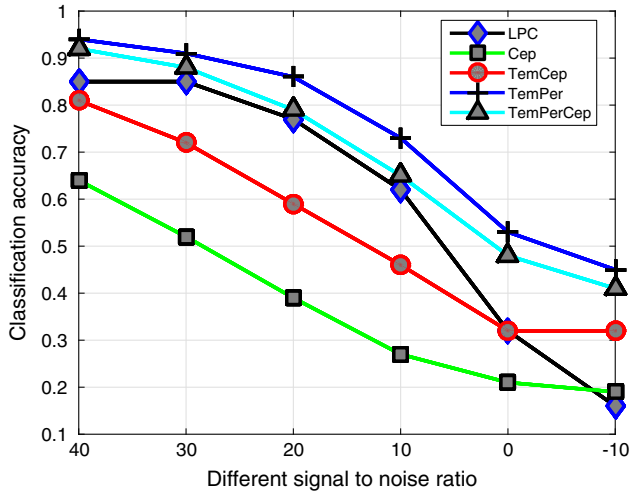


Fig. 9. Sensitivity of different feature representations for different levels of noise contamination.

sensitive to background noise, compared with other feature representations. Comparing *TemCep* with *TemPer*, it can be observed that perceptual features have a better anti-noise ability than the cepstral feature. It is also found that LPC has a good classification performance when SNR is larger than 10 dB, but the classification accuracy quickly decreases when SNR is smaller than 10 dB.

4. Discussion

Table 2 shows the classification performance of previous methods. Since previous studies often used different datasets to perform the classification task, we implement all those features and apply them to the dataset with the same classifier (SVM). Compared with those previous methods, this proposed enhanced feature representation significantly outperforms other methods. Therefore, it can be concluded that our feature representation can effectively characterise different frog calls. From the Table 2, we can also observe that MFCCs is the most popular feature that has been used for frog call classification. Among all used machine learning algorithms, SVM shows the superior performance and is widely used for the classification task. It can also be found that the classification accuracy of *TemPerCep* does not show significant improvement when compared with MFCCs. However, combining temporal and perceptual features with cepstral features greatly improves the anti-noise ability of MFCCs.

5. Conclusion and future work

In this paper, we proposed a novel enhanced feature representation to classify frog calls with various machine learning algorithms. After segmenting continuous recordings into individual syllables, a variety of acoustic features are extracted from each syllable. Then, different features are fused to form different feature representations. Finally, various machine learning algorithms are used to classify frog calls with different feature representations. Our proposed enhanced feature representation shows the best classification accuracy and has good anti-noise ability. Meanwhile, the SVM and RF outperform the traditional LDA and K-NN classifiers. Therefore, it is suitable to combine *TemPerCep* with SVM or RF to build a frog call classification system. Ecologists can apply the proposed classification system to long-term frog recordings. Then, the long-term change of frog species richness can be reflected by the classification results.

Table 2
Comparison with previous used feature representations.

Refs.	Feature	Accuracy (%)
[24,11]	LPCs	93.5
[19,14,17,13]	MFCCs	94.9
[9]	Spectral centroid, Shannon entropy, Rényi entropy	75.6
[14]	Syllable duration, dominant frequency, oscillation rate, frequency modulation, energy modulation	92.3
[12]	Spectral centroid, signal bandwidth, spectral roll-off, threshold-crossing rate, spectral flatness, and average energy	95.8
Our feature representation		
TemPerCep		99.1

In the future, since the MFCCs feature shows a good classification performance, but a bad anti-noise ability, we can modify MFCCs to improve the anti-noise ability. After transforming frog audio data into its spectrogram representation, the visual inspection motivates us to use image processing algorithms for frog calls. All the feature used in this study are calculated based on STFT. It might be worth investigating other techniques, such as Hilbert Huang transform and wavelet transform, for frog call analysis, because those techniques have been successfully used for studying sounds [25]. Also, a wider variety of frog audio data from different geographical and environmental conditions will be tested in the future experiments.

Acknowledgements

Thanks to the QUT Eco-acoustics Research Group for providing the datasets used in this experiment, as well as to the support from the Wet Tropics Management Authority, Queensland, Australia. Thanks to the anonymous reviewers for their careful work and thoughtful suggestions that have helped improve this paper substantially.

All funding for this research was provided by the Queensland University of Technology and the China Scholarship Council (CSC).

References

[1] Wimmer J, Towsey M, Planitz B, Williamson I, Roe P. Analysing environmental acoustic data through collaboration and automation. Future Gener Comput Syst 2013;29:560–8.
[2] Wells KD. The ecology and behavior of amphibians. University of Chicago Press; 2010.
[3] Gingras B, Fitch WT. A three-parameter model for classifying anurans into four genera based on advertisement calls. J Acoust Soc Am 2013;133:547–59.
[4] Xie J, Towsey M, Yasumiba K, Zhang J, Roe P. Detection of anuran calling activity in long field recordings for bio-acoustic monitoring. In: 2015 IEEE tenth international conference on intelligent sensors, sensor networks and information processing, Singapore.
[5] Grigg G, Taylor A, Mc Callum H, Watson G. Monitoring frog communities: an application of machine learning. In: Proceedings of eighth innovative applications of artificial intelligence conference, Portland Oregon. p. 1564–9.
[6] Lee C-H, Chou C-H, Han C-C, Huang R-Z. Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis. Pattern Recogn Lett 2006;27:93–101.
[7] Huang C-J, Yang Y-J, Yang D-X, Chen Y-J. Frog classification using machine learning techniques. Expert Syst Appl 2009;36:3737–43.
[8] Acevedo MA, Corrada-Bravo CJ, Corrada-Bravo H, Villanueva-Rivera LJ, Aide TM. Automated classification of bird and amphibian calls using machine learning: a comparison of methods. Ecol Inform 2009;4:206–14.
[9] Han NC, Muniandy SV, Dayou J. Acoustic classification of australian anurans based on hybrid spectral-entropy approach. Appl Acoust 2011;72:639–45.
[10] Chen W-P, Chen S-S, Lin C-C, Chen Y-Z, Lin W-C. Automatic recognition of frog calls using a multi-stage average spectrum. Comput Math Appl 2012;64:1270–81.
[11] Yuan CLT, Ramli DA. Frog sound identification system for frog species recognition. In: Context-aware systems and applications. Springer; 2012. p. 41–50.

[12] Huang C-J, Chen Y-J, Chen H-M, Jian J-J, Tseng S-C, Yang Y-J, et al. Intelligent feature extraction and classification of anuran vocalizations. Appl Soft Comput 2014;19:1–7.
[13] Bedoya C, Isaza C, Daza JM, López JD. Automatic recognition of anuran species based on syllable identification. Ecol Inform 2014;24:200–9.
[14] Xie J, Towsey M, Truskinger A, Eichinski P, Zhang J, Roe P. Acoustic classification of australian anurans using syllable features. In: 2015 IEEE tenth international conference on intelligent sensors, sensor networks and information processing, Singapore, Singapore.
[15] Stewart D. Australian frog calls: subtropical east. Audio CD; 1999.
[16] Harma A. Automatic identification of bird species based on sinusoidal modeling of syllables. 2003 IEEE international conference on acoustics, speech, and signal processing, 2003. Proceedings. (ICASSP'03), vol. 5. IEEE; 2003. p. V–545.
[17] Jaafar H, Ramli DA. Automatic syllables segmentation for frog identification system. In: IEEE 9th international colloquium on signal processing and its applications (CSPA). IEEE; 2013. p. 224–8.
[18] Itakura F. Line spectrum representation of linear predictor coefficients of speech signals. J Acoust Soc Am 1975;57:S35.
[19] Lee C-H, Chou C-H, Han C-C, Huang R-Z. Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis. Pattern Recogn Lett 2006;27:93–101.
[20] Han W, Chan C-F, Choy C-S, Pun K-P. An efficient MFCC extraction method in speech recognition. In: Proceedings IEEE international symposium on circuits and systems, 2006. ISCAS 2006. IEEE; 2006. p. 4.
[21] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.
[22] Chang C-C, Lin C-J. Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2011;2:27.
[23] Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 2005;21:2185–90.
[24] Juan Mayor LMM. Frogs species classification using LPC and classification algorithms on wireless sensor network platform. In: XVII General Assembly, Ibero-American Conference on Trends in Engineering Education and Collaboration, ISTEC.
[25] Wang S, Zeng X. Robust underwater noise targets classification using auditory inspired time–frequency analysis. Appl Acoust 2014;78:68–76.