Queensland University of Technology
Brisbane Australia

This is the author's version of a work that was submitted/accepted for publication in the following source:

Xie, Jie, Towsey, Michael, Truskinger, Anthony, Eichinski, Philip, Zhang, Jinglan, & Roe, Paul
(2015)
Acoustic classification of Australian anurans using syllable features. In *10th International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2015)*, 7-9 April 2015, Singapore.

This file was downloaded from: http://eprints.qut.edu.au/89673/

© Copyright 2015 IEEE

Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Notice: *Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source:*

http://doi.org/10.1109/ISSNIP.2015.7106924

Acoustic classification of Australian anurans using syllable features

Authors

ABSTRACT—**Acoustic classification of anurans (frogs) has received increasing attention for its promising application in biological and environment studies. In this study, a novel feature extraction method for frog call classification is presented based on the analysis of spectrograms. The frog calls are first automatically segmented into syllables. Then, spectral peak tracks are extracted to separate desired signal (frog calls) from background noise. The spectral peak tracks are used to extract various syllable features, including: syllable duration, dominant frequency, oscillation rate, frequency modulation, and energy modulation. Finally, a k-nearest neighbor classifier is used for classifying frog calls based on the results of principal component analysis. The experiment results show that syllable features can achieve an average classification accuracy of 90.5% which outperforms Mel-frequency cepstral coefficients features (79.0%).**

Keywords—audio classification; syllable feature; principal component analysis; k nearest neighbour; spectral peak track

I. Introduction

Acoustic sensor networks are a well-established and widely deployed method of collecting acoustic data for monitoring animals [1]. The traditional field survey methods that require ecologists to physically visit sites for collecting bio-diversity data are both time-consuming and costly. Comparatively, sensors can record acoustic data automatically, objectively, and continuously for long durations. However, analyzing the large amount of collected data manually is very time-consuming. Developing semi-automatic or automatic methods for classifying collected acoustic data by sensors is thus in high demand and has attracted a lot of research [2-7].

Prior call classification research typically adopts the following structure : (1) pre-processing, (2) segmentation, (3) feature extraction, (4) classification [2]. Taylor et al. proposed a system for identifying 22 frog species recorded in northern Australia based on peak values (intensity of spectrogram) [3]. Huang et al. [4] extracted the *spectral centroid*, *signal bandwidth* and *threshold crossing rate* and used these features with k nearest neighbor (k-NN) and support vector machine (SVM) classifiers to classify frog calls. Dayou et al. [5] developed a method based on entropy to recognize frog calls. *Shannon entropy*, *Renyi entropy* and *Tsallis entropy* were trialed as inputs to a k-NN classifier for recognition. A multi-stage average spectrum was proposed by Chen et al [6]. *Syllable length* was first used for the pre-classification. Then the *multi-stage average spectrum* was extracted for the classification. Chen et al. [7] described the semi-automatic bird call classification method based on spectral peak tracks. A set of spectral features were derived by time-varying analysis of the recorded bird vocalizations for classification. Tyagi et al. [8] proposed the *spectral ensemble average voice* to do bird recognition. Then, *dynamic time warping* was combined to improve the recognition accuracy. Lee et al. [9] introduced a recognition method based on the analysis of spectrogram to detect each syllable. Mel-frequency cepstral coefficients features (MFCCs) of each frame were defined as features, and *linear discriminant analysis* was used for classifying 30 kinds of frog calls and 19 kinds of cricket calls.

Most prior work often reports high accuracy rates for recognition and classification. However, most features used in the prior work are based on only either on only frequency domain or time domain information. However, a combination of the two will be able to discriminate between a wider variety of species that may share similar characteristics in either time or frequency information but not both. This research presents a novel feature extraction method for frog call classification which includes both time and frequency domain information.

After segmenting input frog calls into syllables, the spectral peak track (SPT) algorithm is applied for locating the frog call frequency boundary. Then, the syllable features are extracted from the SPT results. Principal component analysis (PCA) is applied to decorrelate the syllable features and to reduce the dimensionality. Finally, a k-NN classifier is used to classify the frog calls. The proposed syllable features achieve higher classification accuracy (90.5%) than MFCCs (79.0%).

The rest of this paper is organized as follows: In section II, we describe the method for frog call classification, which includes data set acquisition, syllable segmentation, feature extraction, PCA and classification. Section III reports experiment results. Section IV presents conclusion and future work.

II. Method

Our frog call classification method consists of five steps: data set acquisition, syllable segmentation, feature extraction, PCA and classification (Fig.1). Detailed information of each step is shown in following sections.

A. Data set acquisition

In this study, 16 frog species which are widespread in Queensland, Australia are selected for experiments (Table I). All the recordings are obtained from David Stewart [10], and has a sample rate of 44.1 kHz. All recordings were all

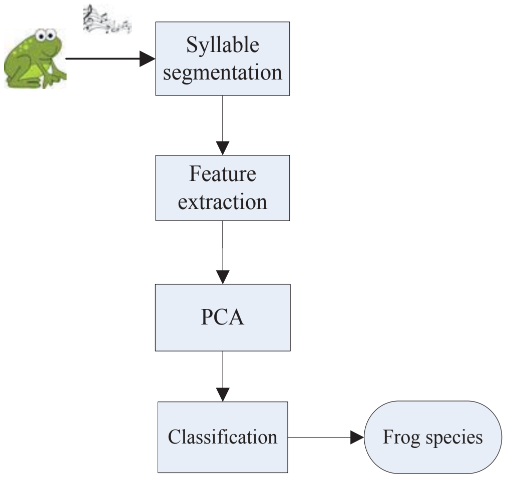


Fig. 1. Flowchart of frog call classification system

mixed-down to mono. 50% dataset was used as training data, and the rest for testing.

B. Syllable segmentation

One syllable is a continuous anuran vocalization emitted from an individual, which is one elementary acoustic unit for classification. In this study, audio data is automatically segmented into a set of syllables using the method proposed by Härmä [11] which is described as follows:

Step 1: Compute the spectrogram (Fig.4) of audio data using a short-time Fourier transform (STFT) (Hamming window, $\text{size} = 512\ \text{samples}, \text{overlap} = 25\%$). We denote the spectrogram as a matrix $\mathrm{S(f,t)}$, where f represents the frequency index and t is the time index.

Step 2: Smooth the original spectrogram using Gaussian filter (5×5) to remove small gaps within syllables. This step is the only deviation from the original technique by Härmä.

Step 3: Find f_n and t_n that $|\mathrm{S}(\mathrm{f}_n, \mathrm{t}_n)| \geq |\mathrm{S(f,t)}|$ for every pair of $\mathrm{(f,t)}$, and set the position of the n^{th}syllable to be t_n.

Step 4: Compute the amplitude of the first frame $\mathrm{A}_n(0) = 20log_{10}|(\mathrm{f}_n, \mathrm{t}_n)|$ decibel (dB). If $\mathrm{A}_n(0) < \mathrm{A}_0(0) - \beta$, stop the segmentation process, where β is the stopping criteria and its default value is 18 dB. If stopped, it means that the amplitude of the n^{th} syllable is too small and hence no more syllables need to be extracted.

Step 5: Start from t_n, trace the maximal peak of $|\mathrm{S(f,t)}|$ for $\mathrm{t} < \mathrm{t}_n$ until $\mathrm{A}_n(t_n - t) = \mathrm{A}_n(0) - \beta$, where $\mathrm{A}_n(t_n - t) = 20\log_{10}|S(f,t)|$. Next, trace the maximal peak of $|\mathrm{S(f,t)}|$ for $\mathrm{t} > \mathrm{t}_n$ until $\mathrm{A}_n(t - t_n) = \mathrm{A}_n(0) - \beta$, where $\mathrm{A}_n(t - t_n) = 20\log_{10}|S(f,t)|$. Hence, the starting and stopping time of the n^{th} syllable are determined as $t_n - t_s$ and $t_n + t_e$.

Step 6: Save the amplitude trajectories to the n^{th} syllable and set $\mathrm{S(f,[\,t_n - t_s, \dots, t_n + t_e]) = 0}$. Repeat steps 3-5. Fig.2. shows the spectrogram of *Mixophyes fasciolatus* (Great Barred Frog) and plots the segmentation result on the waveform for better display.

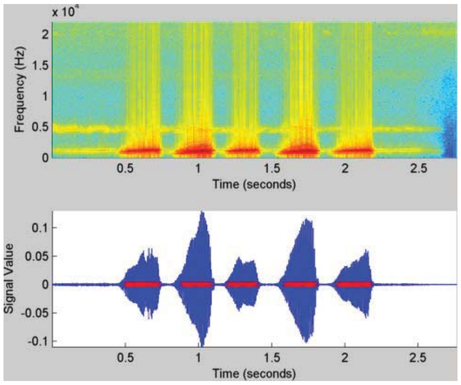


Fig.2. Segmantaion result marked with red line

C. Feature extraction

Five features are extracted from each syllable for frog call classification. They are syllable duration, dominant frequency, oscillation rate, frequency modulation, and energy modulation. MFCCs are used as baseline for comparison.

- Extraction of syllable features

Syllable features are extracted from spectral peak tracks (SPTs), which in turn, isolate the desired signal within the syllable. The SPT method has been used for bird calls in previous research [7]. Here, it is adapted for analyzing frog calls. The SPT method works by matching peaks in the spectrogram from one time frame to the next to produce a

TABLE I. SUMMARY OF THE FROG SCIENTIFC NAME ,COMMON NAME AND CORRESPONDING CODE

No.	Scientific name	Total syllable	Common name	Code
1	*Assa darlingtoni*	36	Pouched frog	ADI
2	*Crinia parinsignifera*	40	Eastern Sign-bearing Frog	CPA
3	*Litoria caerulea*	72	White's tree frog	LCA
4	*Litoria chloris*	26	Red-eyed tree frog	LCS
5	*Litoria latopalmata*	169	Broad-palmed frog	LLA
6	*Litoria nasuta*	60	Striped rocket frog	LNA
7	*Litoria revelata*	151	Whirring Tree Frog	LEA
8	*Litoria rubella*	37	Desert tree frog	LRA
9	*Litoria verreauxii*	28	Verreauxii's tree frog	LVV
10	*Litoria tyleri*	117	Tyler's tree frog	LTI
11	*Limnodynastes tasmaniensis*	14	spotted grass frog	LTS
12	*Limnodynastes terraereginae*	44	Northern banjo frog	LTE
13	*Mixophyes fasciolatus*	28	Great Barred Frog	MFS
14	*Philoria kundagungan*	22	Mountain Frog	PKN
15	*Uperoleia fusca*	32	Dusky Toadlet	UFA
16	*Uperoleia laevigata*	24	Smooth Toadlet	ULA

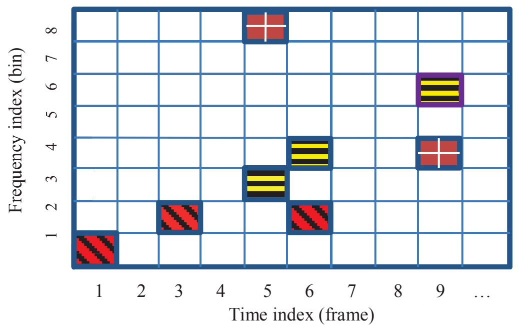


Fig. 3. Peaks in the spectrogram. The red ractangle represets extracted peaks, the yellow rectangle represets predicted postion. The red rectangles with cross do not satify conditions (1) or (2).

connected sequence which shows the amplitude and frequency trajectory of the underlying events [12]. The major steps for extracting SPTs are described as follows: **Step 1:** For input frog calls, the spectrogram is generated using a STFT (Hamming window, size = 128 samples, overlap = 85%).

Step 2: For each frame, the maximum intensity is selected with a minimum required value of 3 dB. This can result in not all time frames containing peaks. P_n denotes a peak with n representing the peak index (not the time frame it is contained in). The 3 dB threshold is chosen empirically.

Step 3: Next, the SPT algorithm is applied to the extracted peaks from step 2. Before describing the algorithm, the following parameters need to be defined (these values are manually tuned for the classification frog species): (1) maximum time domain interval for connecting peaks (1.28 ms), (2) maximum time domain interval for discarding the peaks (4.27 ms), (3) minimum track length (8.54 ms), (4) maximum frequency domain interval (516 Hz), (5) density (the ratio between the number of peaks and the length of the SPT) threshold (0.8). The time domain and frequency domain intervals between two successive peaks are first calculated. If conditions (1) and (4) are satisfied, then a SPT (SPT_1) is generated. For extending SPT_1, linear regression is used to predict next likely continuation of the track. Based on peaks $p_1(t_1, f_1)$ and $p_2(t_2, f_2)$, β and ε in equation (1) can be solved.

$$f = \beta t + \varepsilon \qquad (1)$$

Then, one by one, the predicted peak $\tilde{p}_n$ of the following frame t_n can be calculated, shown as the yellow rectangle in Fig. 3. If the time interval between p_2 and p_n does not satisfy condition (1), p_n will not be added to SPT_1, and we move to the next peak p_{n+1}. Otherwise, we calculate the frequency interval between $\tilde{p}_n$ and p_n. If condition (4) is satisfied, then p_n will be added to SPT_1. After each peak is added, linear regression is repeated to recalculate the next predicted peak using at most the last 10 included peaks. This iterative process continues until condition (2) is no longer satisfied. Once SPT_1 stops growing, the length and the density of SPT are then calculated. If the results satisfy condition (3) and (5), SPT_1 will be stored. Each SPT_n is stored as: start time t_s , stop time t_e and frequency bin index of each of the peaks within the track f_t, $t_s \le t \le t_e$. The result of the SPT algorithm is shown in Fig.4.

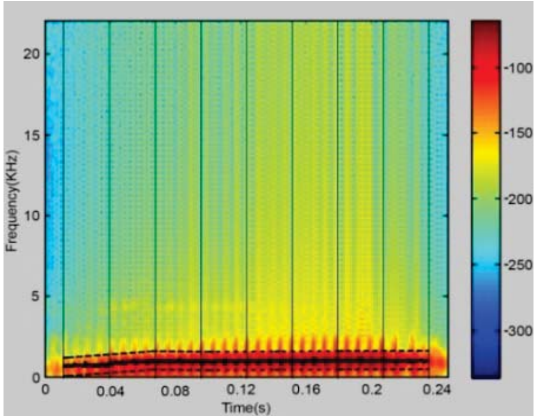


Fig.4. SPT (solid horizontal black line) on the spectrogram, the dash horizontal black lines represent the frequency boundary of the SPT. We calculate 19 variables for each call within the frequecy boundary: syllable duration, dominant frequency, oscillation rate, frequency modulation and energy modulation

Each syllable is represented as a single SPT, the syllable features are then extracted from the results of the SPT algorithm.

a) Syllable duration (seconds): the syllable duration (D) is directly obtained from the bounds (time domain) of the segmentation result.

$$D = (t_e - t_s)/r_x \qquad (2)$$

where r_x is the x-axis resolution.

b) Dominant frequency (Hz): the dominant frequency ($\hat{f}$) is calculated by averaging the frequency of all peaks within one SPT.

$$\hat{f} = \sum_{t=t_s}^{t_e} f_t/(t_e - t_s + 1) * r_y \qquad (3)$$

where r_y is the y-axis resolution, f_t is the frequency bin index of peak t.

c) Oscillation rate (Hz): the oscillation rate (O_r) is the number of pulses within one second. The algorithm for extracting oscillation rate is a modified form of Bardeli's [13], which is described as follows:

1. Calculate dominant frequency bin ($\hat{f}$) of the SPT and define the frequency domain boundary as

   $[l, h] = [\max(\hat{f} - 5, 1), \hat{f} + 5]$. Here, the value 5 is determined empirically. The power within the boundary is calculated as

   $$P_t = \sum_{f=l}^{h} S(t,f) * (S(t,f) - S(t+1,f))^2 \qquad (4)$$

2. Normalize P_t to [0,1] and discard the first 20% and last 20% part of the signal as the signal towards the start and end of the syllables is often less clear.

3. Calculate the autocorrelation with the length of the selected vector and the result is represented by P_t^{ACF}.

4. Subtract the mean from P_t^{ACF}

$$\nabla P_t^{ACF} = P_t^{ACF} - mean(P_t^{ACF}) \tag{5}$$

Then, a discrete cosine transform (DCT) is applied to ∇P_t^{ACF} for isolating different frequency components. The DCT of the auto-correlated power is defined as

$$\mathrm{P}_d(\mathrm{k}) = \mathrm{w}(\mathrm{k}) \sum_{j=1}^{N} \nabla P_t^{ACF} cos \frac{\pi(2j-1)(k-1)}{2N}, \mathrm{k} = 1, \dots, \mathrm{N} \tag{6}$$

Here, the DCT length (N) is 0.2s.

$$\mathrm{w}(\mathrm{k}) = \begin{cases} \frac{1}{\sqrt{N}}, k = 1 \\ \sqrt{\frac{2}{N}}, 2 \leq k \leq N \end{cases} \tag{7}$$

5. Set $\mathrm{P}_d(\mathrm{k}), k = 1, \dots, 5$ to zero for removing low frequency oscillation from consideration. The oscillation rate is then calculated using the location of the highest power L_p:

$$O_r = \frac{L_p}{D} * 0.5/r_x \tag{8}$$

d) Frequency modulation (Hz)

Frequency modulation ($\mathrm{F}_m, m = 1 \dots 8$) means the difference between the dominant frequency bin ($\hat{\mathrm{f}}$) and averaged frequency bin $\bar{f}_l, l = [1,2, \dots, 8]$ of eight equal segments of the SPT.

$$\mathrm{F}_m = [\bar{f}_1 - \hat{\mathrm{f}}, \dots \bar{f}_8 - \hat{\mathrm{f}}] * \mathrm{r}_y \tag{9}$$

e) Energy modulation

Energy modulation ($\bar{E}_m, m = 1 \dots 8$) means the averaged energy of eight equal parts of the SPT. Let $A(n)$ denote the amplitude value of the SPT, where n is the frame index. First, a Hilbert transform is applied to A(n) as follows:

$$A_h(n) = Hilbert(A(n)) \tag{10}$$

where $A_h(n)$ is the complex sequence named as the analytic signal of $A(n)$.

Then, the absolute part of $A_h(n)$ is extracted to represent the envelope of $A(n)$, and is defined as

$$B(n) = abs(A_h(n)) \tag{11}$$

Lastly, the energy modulation is calculated based on $B_i(n)$. Here, $B_i(n)$ represents one part. The energy modulation is then obtained.

$$\bar{E}_m = [\bar{E}_1, \dots, \bar{E}_8] \tag{12}$$

where $\bar{E}_i = \frac{1}{L} \sum_{n=0}^{L-1} B_i(n)^2$, L is the length of one syllable.

- Extraction of MFCCs

Mel-frequency cepstral coefficients (MFCCs) computed based on short-time analysis are used as the baseline due to the consistency, easy implementation, and reasonable performance [9]. This is also the case for much other prior research. The steps for MFCCs processing are listed as follows:

Step 1: Pre-emphasis.

$$y(n) = s(n) - \alpha s(n-1) \tag{13}$$

where $s(n)$ is input frog call, a typical value for α is 0.95.

Step 2: Framing and windowing.

Each syllable is separated into frames with a length of 512 samples and an overlap of 256 samples. To reduce the discontinuity on both sides of frames, each frame is multiplied by a Hamming window.

$$x(n) = w(n)y(n) \tag{14}$$

where $w(n)$ is the Hamming window function.

$$w(n) = 0.54 - 0.46cos\left(\frac{2n\pi}{L-1}\right), 0 \leq n \leq L-1 \tag{15}$$

Step 3: Spectral analysis.

Compute the discrete Fourier transform (DFT) of each frame of the signal. By considering $\omega = \frac{2\pi k}{N}$, the DFT of each frame of the signal is

$$\mathrm{X}(k) = \sum_{n=0}^{N-1} x(n) e^{-jw}, \mathrm{k} = 0,1, \dots, \mathrm{N}-1 \tag{16}$$

Equation (16) is known as signal spectrum.

Step 4: Band-pass filtering.

The amplitude spectrum is then filtered using a set of triangular band-pass filters.

$$E_j = \sum_{k=0}^{\frac{N}{2}-1} \emptyset_j(k) A_k, 0 \leq j \leq J-1 \tag{17}$$

where J is the number of filters, $\emptyset_j$ is the j^{th} filter, and A_k is the amplitude of $\mathrm{X}(k)$.

$$A_k = |X[k]|^2, 0 \leq k \leq N/2 \tag{18}$$

Step 5: DCT. MFCCs for the i^{th} frame are computed by performing DCT on the logarithm of E_j.

$$C_m^i = \sum_{j=0}^{J-1} \cos\left(m \frac{\pi}{J}(j+0.5)\right) log_{10}(E_j), \ 0 \leq m \leq L-1 \tag{19}$$

where L is the number of MFCCs.

In this study, the filter bank consists of 40 triangular filters, that is $\mathrm{J} = 40$. The length of MFCCs of each frame is 16 ($L = 16$). After calculating MFCCs from each frame, the averaged MFCCs of all frames within one syllable are calculated.

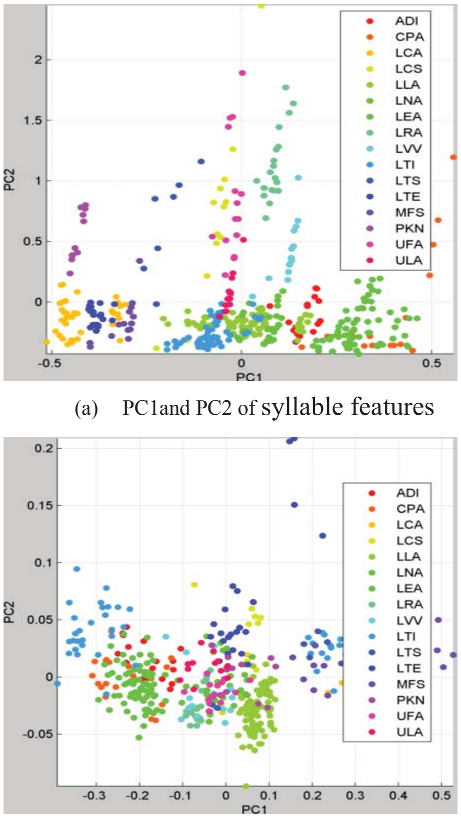
$$f_m = \frac{\sum_{i=1}^{K} C_m^i}{K}, 0 \leq m \leq L-1 \tag{20}$$

where f_m is the m^{th} MFCCs, K is the number of frames within the syllable. In the training phase, the averaging of f_m over all training syllables for the call of the same species is regarded as the m^{th} feature value F_m. A linear normalization process is applied to get the final feature.

$$\tilde{F}_m = \frac{f_m - f_m^{min}}{f_m^{max} - f_m^{min}} \qquad (21)$$

D. Principal component analysis

In this study, dimensionalities of the original syllable features and MFCCs are 19 and 16, respectively. PCA is then applied to decorrelate these feature vectors and reduce their respective dimensionalities. By finding the orthogonal linear combinations (named PC) of the original variables with the largest variance, the dimensionality of the data will be reduced by PCA. The result of PCA is a set of linear combinations of the original features, ordered by their discriminating power. The PCs with the most discriminating power can then be used, and the rest discarded with minimal impact on the results [14]. In this study, the number of PCs for syllable features and MFCCs are 5 and 7 respectively after dimensionality reduction. The distributions of first two PCs for syllable features and MFCCs are shown in Fig.5.



(a) PC1and PC2 of syllable features

(b) PC1 and PC2 of MFCCs feature

Fig.5. Distribution of first two components for (a) syllable features and (b) MFCCs feature

E. Classification

The k-NN classifier is a non-parametric classifier that is appropriate for use because of PCA results [15]. Meanwhile, it has also been widely used for classifying animal calls [4, 5]. Given a set of parameters, a k-NN classifier will find the nearest neighbor among training data by determining the minimum distance between the instances of the testing and training sets. Here, the input parameters for the k-NN classifier are the PCA result of the syllable features and MFCCs features, whose dimensionalities are 5 and 7 respectively. The distance function for the k-NN classifier is Euclidean function and the number of neighbor, k, is 5 which are both selected based on the training data.

III. Experiment

In this experiment, the k-NN classifier is used to learn a model on the training examples with 10-fold cross-validation. Since the k-NN classifier is sensitive to the local structure of the data as well as the initial cluster centroids, we run the k-NN classifier for 10 times based on different initial points. The classification accuracy is defined as follows:

$$\text{Accuracy}(\%) = \frac{N_c}{N_t} \qquad (22)$$

where $\mathrm{N_c}$ is the number of syllables which are correctly classified, and $\mathrm{N_t}$ is the total number of syllables. The k-NN classifier was used with two feature sets: the syllable features and MFCCs. A Gaussian white noise signal, with signal to noise ratio (SNR) of 40 dB, 30 dB, 20 dB, and 10 dB was added to the original audio data for testing the robustness of the syllable features, the results are shown in Fig.7. Table II lists the averaged classification accuracy of syllable features and MFCCs which are 90.5% and 79.0%, respectively. For syllable features, the classification accuracy of *Crinia parinsignifera*, *Limnodynastes tasmaniensis* and *Litoria chloris* is 100%, because the syllable duration, dominant frequency, and oscillation rate of those frog species are stationary and different from

Table II. Comparison of the accuracy of the classifier

Scientific name	**Total syllable**	**Classification accuracy**	
		syllable features	**MFCCs**
Assa darlingtoni	36	94.1%	73.6%
Crinia parinsignifera	40	100%	42.6%
Litoria caerulea	72	89.7%	95.5%
Litoria chloris	26	100%	100%
Litoria latopalmata	169	94.8%	91.2%
Litoria nasuta	60	89.5%	56.6%
Litoria revelata	151	94.7%	80.8%
Litoria rubella	37	79.1%	68.4%
Litoria verreauxii verreauxii	28	94.5%	69.4%
Litoria tyleri	117	91.8%	78.4%
Limnodynastes tasmaniensis	14	100%	100%
Limnodynastes terraereginae	44	94.9%	93.3%
Mixophyes fasciolatus	28	82.9%	100%
Philoria kundagungan	22	94.4%	94.4%
Uperoleia fusca	32	69.8%	53.3%
Uperoleia laevigata	24	77.0%	65.5%
Averaged classification accuracy		90.5%	79.0%

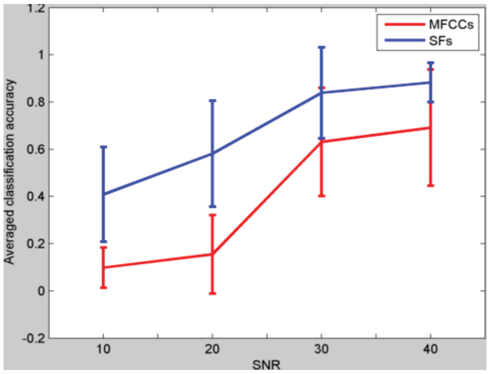


Fig.7. Sensitivity of syllable fatures (SFs) and MFCCs feature for different levels of noise contamination

others. Since the dominant frequency and syllable duration between *Uperoleia fusca* and *Uperoleia laevigata* are similar, the oscillation rate and syllable duration between *Uperoleia fusca* and *Litoria rubella* are similar, the classification accuracy of *Uperoleia fusca*, *Litoria rubella*, and *Uperoleia laevigata* is relatively low. For MFCCs, the classification accuracy of *Limnodynastes tasmaniensis*, *Litoria chloris*, and *Mixophyes fasciolatus* is 100%, because the spectrum distributions of those frog species are different from others. Compared with the MFCCs features, the performance of *Crinia parinsignifera* and *Litoria nasuta* is greatly improved. The results from running the classifier on audio data with artificially added background noise show the ability of our feature extraction method for dealing with background noise.

IV. CONCLUSION

This study presents a novel feature extraction method for classifying frog calls. The audio data is first segmented into syllables. Then, the SPT algorithm is used to isolate frog calls. Syllable features that include syllable duration, dominant frequency, oscillation rate, frequency modulation, and energy modulation are extracted from the boundary of the SPT results for classifying frog calls using a k-NN classifier. The results are promising with an average classification accuracy of 90.5% for syllable features. Future work will include additional experiments that test a wider variety of audio data from different geographical and environment conditions.

REFERENCE

[1] Chesmore, D., Automated bioacoustic identification of species. Anais da Academia Brasileira de Ciências, 2004. 76(2): p. 436-440.

[2] Scott Brandes, T., Automated sound recording and analysis techniques for bird surveys and conservation. Bird Conservation International, 2008. 18(S1): p. S163-S173.

[3] Andrew Taylor , G.W., Gordon Grigg , Hamish Mccallum, Monitoring frog communities: an application of machine learning, in in Proceedings of the 8th Innovative Applications of Artificial Intelligence Conference. Portland Oregeon: AAAI. 1996. p. 1564--1569.

[4] Huang, C.-J., et al., Frog classification using machine learning techniques. Expert Systems with Applications, 2009. 36(2): p. 3737-3743.

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

[6] Chen, W.-P., et al., Automatic recognition of frog calls using a multi-stage average spectrum. Computers & Mathematics with Applications, 2012. 64(5): p. 1270-1281.

[7] Chen, Z. and R.C. Maher, Semi-automatic classification of bird vocalizations using spectral peak tracks. The Journal of the Acoustical Society of America, 2006. 120: p. 2974.

[8] Tyagi, H., et al. Automatic identification of bird calls using spectral ensemble average voice prints. in Proceedings of the 13th European signal processing conference. 2006.

[9] Lee, C.-H., et al., Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis. Pattern Recognition Letters, 2006. 27(2): p. 93-101.

[10] http://www.naturesound.com.au/cd_frogsSE.htm, retrieved on 14th Dec.

[11] Harma, A. Automatic identification of bird species based on sinusoidal modeling of syllables. in Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).

[12] Chatterjee, S. and T.V. Sreenivas, Optimum Transform Domain Split VQ. Signal Processing Letters, IEEE, 2008. 15: p. 285-288.

[13] Bardeli, R., et al., Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recognition Letters, 2010. 31(12): p. 1524-1534.

[14] Jackson, J.E., A user's guide to principal components. Vol. 587. 2005: John Wiley & Sons.

[15] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparamet ricregression". The American Statistician 46 (3): 175–185.