Acoustic classification of Australian frogs for ecosystem surveys

A THESIS SUBMITTED TO
THE SCIENCE AND ENGINEERING FACULTY
OF QUEENSLAND UNIVERSITY OF TECHNOLOGY
IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Jie Xie

School of Electrical Engineering and Computer Science
Science and Engineering Faculty
Queensland University of Technology

April 2016

Copyright in Relation to This Thesis

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature:

Date:

To my family

Abstract

Rapid decreases in frog populations have been spotted from locations over the world, which are regarded as one of the most critical threats to the global biodiversity. Causes of these declines can be summarised as follows: disease, habitat destruction and modification, exploitation, pollution, pesticide use, introduced species, and ultraviolet-B radiation (UV-B). On the one hand frog populations are declining globally, but on the other frogs play an important role in the whole ecosystem. To assess frog populations and optimise the protection policy, monitoring frogs is becoming ever more necessary. Since frogs are much easier to be heard than seen, frog populations are often assessed by their vocalisations. In order to collect frogs' vocalisations, traditional manual methods require ecologists and volunteers to visit the field, which limits the scale for acoustic data collection. In contrast, recent advances in acoustic sensors provide us a novel method to survey vocalising animals such as frogs. Deploying acoustic sensors in the field, they can then automatically collect large volumes of acoustic data at large spatial and temporal scales. After the data is collected by sensors, the raw audio data must be analysed to provide interpretable results. Due to the amount of collected data, enabling automating species identification in acoustic data has become important. Also since the data is collected from the field, the acoustic data tend to be very noisy. Very often the desired signal (frog call) is weak and there are multiple overlapping signals over the frog call. These characteristics pose a big challenge to perform the classification of frog species in acoustic data.

The research presented in this dissertation aims to investigate methods to build a robust and high performance classification system for frog species in acoustic data. By considering the influences for a classification system, two aspects are investigated: feature extraction and classification, which consist of contributions towards three main objectives:

(1) Develop an enhanced feature representation for frog call classification. Classifying frog

calls based on the enhanced temporal, perceptual and cepstral features is proposed. Time-frequency information of frog calls can be effectively represented via the enhanced feature representation. Classification performance of various machine learning techniques is compared with different feature representations. Our proposed enhanced feature representation achieves the best classification accuracy which outperforms the state-of-the-art.

(2) Propose a novel feature representation based on adaptive wavelet packet decomposition. To better capture the frequency domain information of frog calls with a good anti-noise ability, a novel feature representation is proposed named *adaptive frequency scaled wavelet packet decomposition sub-band cepstral coefficients*. Compared with other cepstral coefficients, our proposed feature representation shows the best classification performance and a good anti-noise ability.

(3) Design a robust classification system to study overlapping frog vocalisations. Two classification frameworks are employed to classify overlapping frog vocalisations.

   (a) Multiple-instance Multiple-label (MIML) learning

   To use MIML learning for classifying overlapping frog vocalisations, each individual syllables are first segmented. Then, various features are calculated from each segmented syllable. Next, a bag generator is applied to those extracted features to construct a suitable bag-of-syllable representation. Finally, three MIML learning algorithms are employed for the classification of frog vocalisations: MIML-SVM, MIML-KNN, and MIML-RBF.

   (b) Multiple-label (ML) learning

   As for the ML learning, acoustic features are first calculated without segmentation. Then, ML learning is used to classify simultaneously vocalising frog species using extracted features. Three main ML learning methods are compared: Binary relevance, Classifier Chains, Random k-labelsets, where the base classifier is decision tree. Furthermore, the frog abundance and species richness over three months are calculated based on the results acoustic event detection and ML classification, respectively. Lastly, the correlation analysis between frog calling activity (frog abundance and species richness) and weather variables (mean temperature and rainfall) are studied.

Our proposed approach achieves promising classification results compared to the state-of-the-art. Novel feature representations and classification learning frameworks have different contributions to the performance of the classification system of frog vocalisations. To our knowledge so far, it is the first time that MIML learning and ML learning are employed for automatic classification of overlapping frog vocalisations.

List of Publications

Journal Article

1. **Xie, Jie**, Towsey, Michael, Zhang, Jinglan, and Roe, Paul (2016) Adaptive frequency scaled wavelet packet decomposition for frog call classification. Ecological Informatics, 32, pp. 134-144.

2. Zhang Liang, Towsey Michael, **Xie Jie**, Zhang Jinglan, Roe Paul, Using multi-label classification for acoustic pattern detection and assisting bird species surveys, Applied Acoustics, Volume 110, September 2016, Pages 91-98, ISSN 0003-682X,

Conference Paper

1. **Xie, Jie**, Towsey, Michael, Zhang, Jinglan, and Roe, Paul, Detecting frog calling activity based on acoustic event detection and multi-label learning, International Conference on Computational Science (Accepted)

2. **Xie, Jie**, Towsey, Michael, Zhang, Liang, Zhang, Jinglan, and Roe, Paul, Multiple-Instance Multiple-Label Learning for the Classification of Frog Calls With Acoustic Event Detection, International Conference on Image and Signal Processing (Accepted)

3. **Xie, Jie**, Towsey, Michael, Zhang, Liang, Zhang, Jinglan, and Roe, Paul, Feature Extraction Based on Bandpass Filtering for Frog Call Classification, International Conference on Image and Signal Processing (Accepted)

4. **Xie, Jie**, Towsey, Michael, Truskinger, Anthony, Eichinski, Philip, Zhang, Jinglan, and Roe, Paul (2015) Acoustic classification of Australian anurans using syllable features. In 2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), IEEE, Singapore, pp. 1-6.

5. **Xie, Jie**, Towsey, Michael, Yasumiba, Kiyomi, Zhang, Jinglan, and Roe, Paul (2015) 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 (ISSNIP), IEEE, Singapore, pp. 1-6.

6. **Xie, Jie**, Towsey, Michael, Zhang, Jinglan, and Roe, Paul (2015) Image processing and classification procedure for the analysis of Australian frog vocalisations. InProceedings of the 2nd International Workshop on Environmental Multimedia Retrieval, ACM, Shanghai, China, pp. 15-20.

7. **Xie, Jie**, Towsey, Michael, Zhang, Jinglan, Dong, Xueyan, and Roe, Paul (2015)Application of image processing techniques for frog call classification. In IEEE International Conference on Image Processing (ICIP 2015), 27-30 September 2015, Qubec City, Canada.

8. **Xie, Jie**, Towsey, Michael, Eichinski, Philip, Zhang, Jinglan, and Roe, Paul (2015)Acoustic feature extraction using perceptual wavelet packet decomposition for frog call classification. In 2015 IEEE 11th International Conference on e-Science (e-Science), IEEE, Munich, Germany, pp. 237-242.

9. **Xie, Jie**, Zhang, Jinglan and Roe, Paul, Discovering acoustic feature extraction and selection algorithms for frog vocalization monitoring with machine learning techniques, 2015 Annual Conference of the Ecological Society of Australia. (Abstract accepted for poster presentation)

10. **Xie, Jie**, Zhang, Jinglan, and Roe, Paul (2015) Acoustic features for hierarchical classification of Australian frog calls. In 10th International Conference on Information, Communications and Signal Processing, 2-4 December 2015, Singapore.

11. Dong, Xueyan, **Xie, Jie**, Towsey, Michael, Zhang, Jinglan, and Roe, Paul (2015)Generalised features for bird vocalisation retrieval in acoustic recordings. In IEEE International Workshop on Multimedia Signal Processing, 19-21 October 2015, Xiamen, China.

Keywords

Bio-acoustic monitoring

Environment audio analysis

Frog call classification

Spectrogram analysis

Feature fusion

Wavelet packet decomposition

Multiple-instance multiple-label learning

Multiple-label learning

Acknowledgments

First, I would like to express my sincere gratitude and thanks to Dr Jinglan Zhang (principal supervisor). I want to thank Jinglan for giving me an opportunity to study in Australia. During the whole PhD study, I learnt so much from her about passion for work and high motivation, which will benefit me throughout my life. I would also like to express my gratitude to Professor Paul Roe (associate supervisor), for his consistent instructions and supports through the last three years. Especially, Paul helped me pay the registration fee for several conferences, which makes our accepted paper successfully included in the conference proceedings.

I would also like to thank Dr Michael Towsey for his provision of consistent guidelines, discussions, and encouragement in the first year of my Phd study. Michael's attitude towards the scientific research keeps motivating me go deep into the research.

I want to thank Professor Vinod Chandran for his support in writing my confirmation report and this thesis. Vinod's strong background knowledge in signal processing greatly help me improve my understanding of this research.

I would also like to express my gratefulness to my family, especially my grandparents, parents and my wife. They always support my oversea study silently and firmly. Without their support, I could not pay full attention to PhD study and complete this thesis. My sincere thanks also go to all the friends for their love, attention and continuous concern about my PhD study.

I also want to thank the China Scholarship Council (CSC), Queensland University of Technology and the Wet Tropics management authority for their financial support.

Table of Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

DFT	Discrete Fourier Transform
DCT	Discrete Cosine Transform
STFT	Short-Time Fourier Transform
LPCs	Linear Predictive Coding
MFCCs	Mel-Frequency Cepstral Coefficients
LDA	Linear Discriminant Analysis
K-NN	K-Nearest Neighbour
SVM	Support Vector Machine
ANN	Artificial Neural Network
RF	Random Forest
AED	Acoustic Event Detection
WPD	Wavelet Packet Decomposition
MIML	Multiple-Instance Multiple-Label
ML	Multiple-Label

Chapter 1

Introduction

1.1 Motivation and background

During the past decades, rapid decreases in frog populations have been spotted from locations over the world, which are regarded as one of the most critical threats to the global biodiversity. Many environment problems are regarded as the reasons for these declines: disease, habitat destruction and modification, exploitation, pollution, pesticide use, introduced species, and ultraviolet-B radiation (UV-B). On one hand frog populations are rapidly worldwide declining, and on the other frogs are greatly important to the global ecosystem.

(1) Frogs are integral part of the food web

(2) Frogs are often used as the environment indicators

(3) Frogs are important in medical research that benefits humans

For those aforementioned reasons, increasing frog populations and optimising the protection policy necessitates monitoring of frogs. Frogs are often much easier to be heard than to be seen (Figure. 1.1). Also frog vocalizations are often employed for most communication, which offer a possible way to study and evaluate frog populations by detecting species-specific calls [Dorcas et al., 2009]. Therefore, frogs are often monitored via their vocalisations. Traditional manual monitoring methods require ecologists and volunteers to spend extensive time in the field for collecting acoustic data. Although traditional methods can provide an accurate measure of daytime species and richness, it has a limitation in monitoring frog populations over large spatial and temporal scales. To address this limitation, recent advances in acoustic sensors provide a

way to automatically survey vocal animals (such as frogs). Deploying acoustic sensors in the field, frog vocalisations can then be automatically collected. Compared with the manual point-counting method, sensors can greatly extend the survey into larger spatial and temporal scales, and generate large volumes of acoustic data that needs to be analysed. Consequently, enabling automatic species identification in acoustic data has become important. However, because the recordings are automatically collected from the field, the audio data tends to be very noisy. Very often the desired signal (frog call) is weak, and there are multiple overlapping signals over the frog call. Furthermore, different frog species tend to call together to make chorus. All those characteristics pose a big challenge to automatic frog vocalisations survey.

Figure 1.1: Photos of frogs to indicate that frogs are difficult to be seen in the field

1.2 Basic concepts

1.2.1 Environment audio data

The audio data used in this study is mainly derived from two sources: David Stewart's CD [Stewart, 1999] and recordings collected by James Cook University (JCU) [1]. David Stewart's CD is employed for the preliminary testing, and used for the experiments in chapters 4 and 5. Recordings collected by JCU are used for chapters 6 and 7. The reasons for using those two datasets are list as follows:

- Since almost all prior work studied frog recordings with an assumption that only one frog species exists in each individual frog recording, the experiments in Chapter 4 and 5 aim

[1] All the recordings can be obtained from our group website: https://www.ecosounds.org/

to develop a state-of-the-art frog call classification system under this assumption.

- Chapters 6 and 7 focus on the study of recordings including multiple simultaneously vocalising frog species, which is the real situation for most environmental recordings.

Compared with audio data collected in the laboratories and quiet places (such as David Stewarts CD), environmental recordings are normally collected under unconstrained noisy conditions (such as JCU recordings). Consequently, the noise and variability issues need to be considered when dealing with environmental audio data. For the background noise, there are a wide variety of non-biological noises and a variety of animal sounds in the environmental recordings. These non-biological noises often come from different sources: rain, wind, human activities (e.g. traffic noise). Besides non-biological noises, many competitive animal sounds (e.g. birds when we are interested in frogs) are also recorded in the environmental recordings. In the case of variability, it is produced in many aspects: call structure between species, population of one specific species, time and season changes. All those noises and variabilities make it a challenge to develop a robust frog call classification system.

1.2.2 Audio data analysis

Audio data is usually considered as a mono-dimensional signal. To ease the tasks of understanding, comparison, modification, and resynthesise of signals [Rocchesso, 2003], audio data analysis is often developed to find the major features representing the time-varying audio data. Many application areas of audio data analysis have been identified: speech processing, mechanical signal processing, bioacoustics analysis, etc. Two most important audio data analysis techniques are Short-time Fourier Transform (STFT) and Linear Predictive Coding (LPC). STFT is a Fourier-related transform, which determines the sinusoidal frequency and phase content of local sections of a signal as it changes over the time [Allen, 1997]. LPC is mostly used to represent the spectral envelope of an audio data based on the information of a linear predictive model [Deng and O'Shaughnessy, 2003]. An example of a spectrogram of frog calls derived from a field recording is shown in Figure. 1.2.

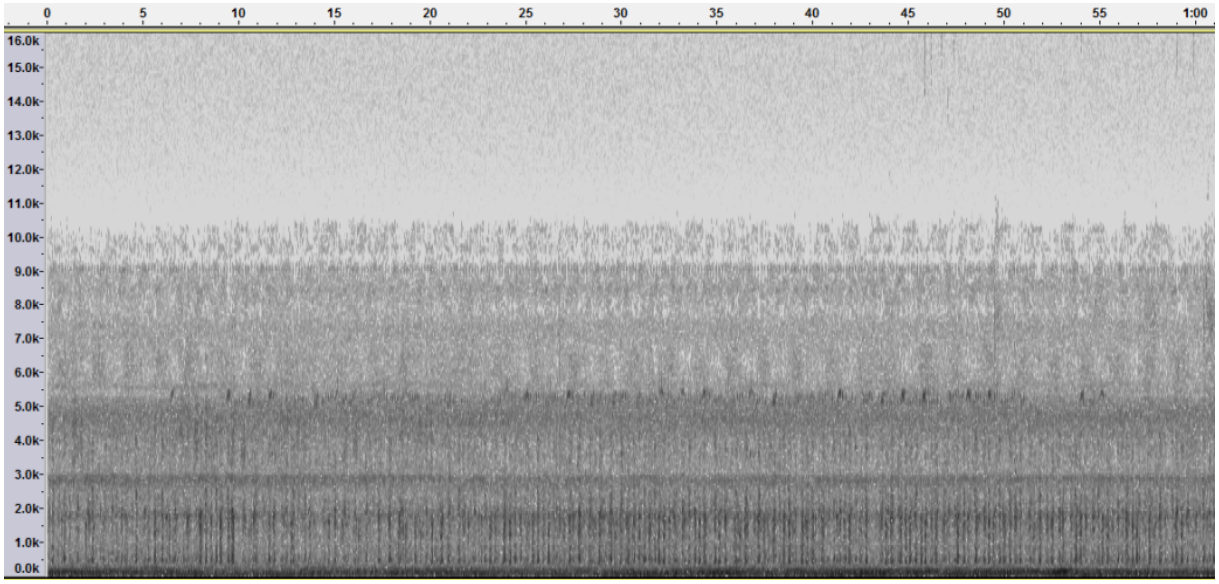


Figure 1.2: An example of spectrogram of environmental recording. The x-axis is time (seconds); the y-axis is frequency (kHz). The spectrogram is generated from a one-minute recording collected in Townsville, Queensland on around 11.50 pm February 03 2013; the frog species in this recording is *Litoria caerulea*

1.2.3 Frog call structure

Spectrogram (also called sonogram) is a widely used tool for most bioacoustics analysis for its flexible implementation and good applicability. Compared with the hierarchical structure of bird calls, frog calls have a relatively simple call structure [Somervuo et al., 2006]. The frog vocalisation structure mainly has two ingredients: call and syllable. A frog call is normally made up of several frog syllables (Figure. 1.3).

One syllable is basically a sound that a frog produces with a single blow of air from the lungs [Huang et al., 2009]. For frog call classification, an elementary unit is one syllable. To get an intuitive sense of frog call structure, examples of different frog species in both waveform , spectrogram, and signal-to-noise ratio (SNR) are shown in Table 1.1 and Table 1.2. For the waveform, x-axis and y-axis represent time and amplitude scales, respectively. The x-axis and y-axis of the spectrogram represent the time and frequency scales, respectively. The grey scale represents the acoustic intensity. Six frog species, which are widely distributed in Queensland, Australia, are selected from David Stewart's CD to generate waveform and spectrogram [Stewart, 1999]. For JCU recordings, eight frog species are selected. The SNR is calculated as follows:

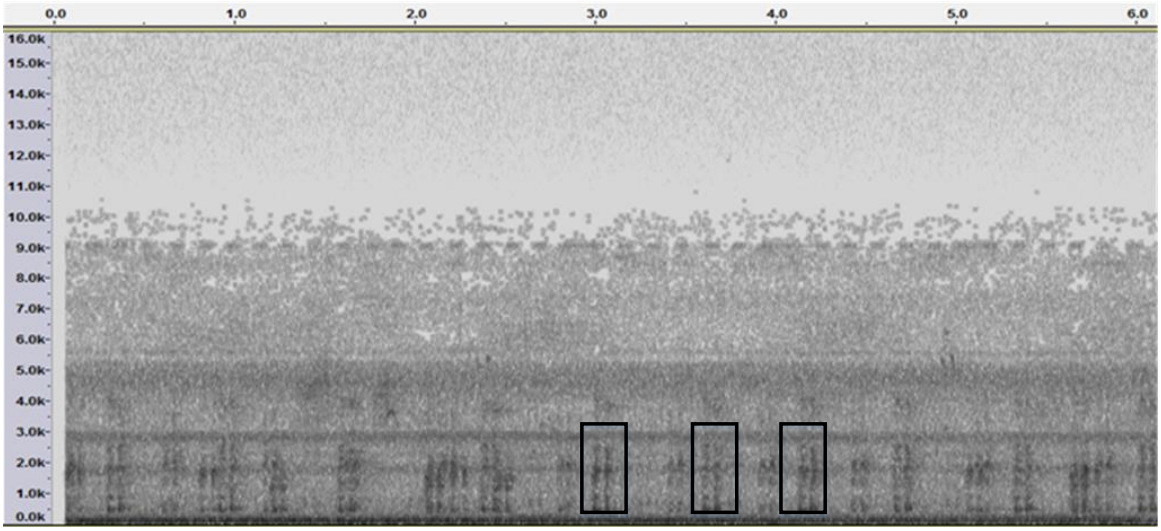


Figure 1.3: Spectrogram of *Litoria caerulea*, three syllable of *Litoria caerulea* are annotated with one black rectangle, respectively.

$$SNR = 10 * log_{10}(\frac{\sum_{i=m}^{m+L} S_i^2}{\sum_{j=n}^{n+L} N_j^2}) \tag{1.1}$$

where L is the length of the signal and noise used for calculating SNR, and set at 6000 samples here, n and m are manual selected start location in the waveform for noise and signal, respectively.

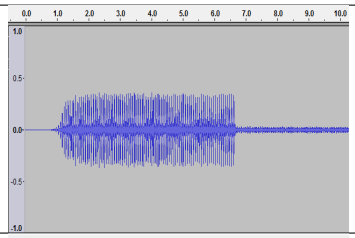
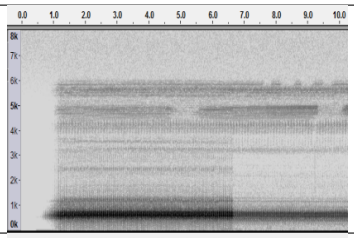
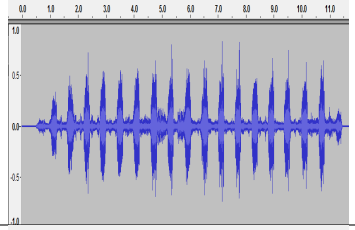
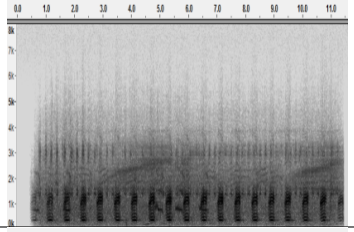
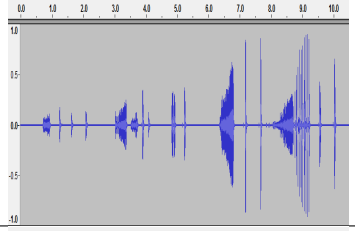
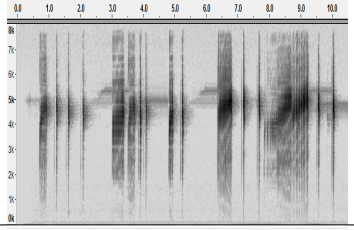
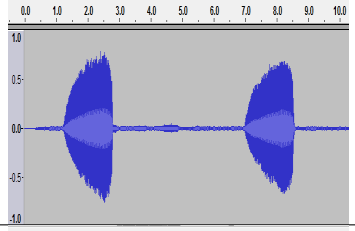
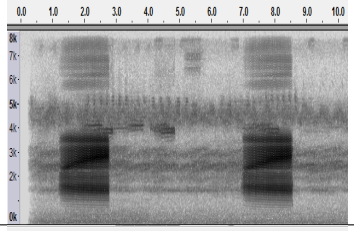
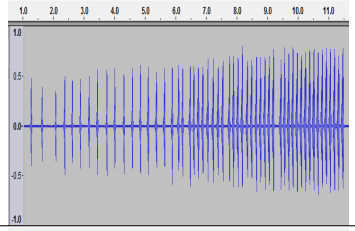
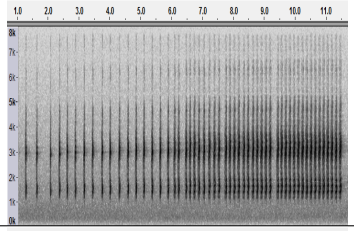
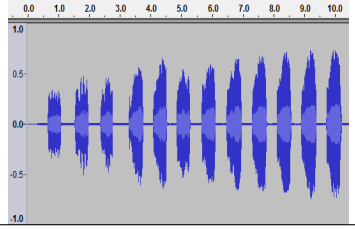
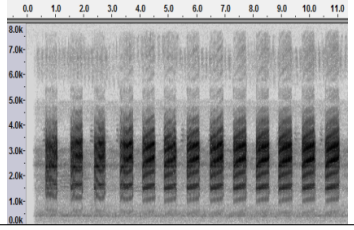
The parameters of those frog species are shown in Table 1.3.

Since the power of signal and background noise in this study vary from recordings to recordings and calls to calls within the recording, Table 1.4 and Table 1.5 calculate the confidence intervals of high and low SNR recordings for the power of signal and noise. The calculation of confidence interval is defined as follows:

$$CI = \mu \pm Z * \frac{\sigma}{\sqrt{L}} \tag{1.2}$$

where μ and σ are the mean and standard deviation, respectively, Z is the upper $\frac{(1-C)}{2}$ critical value for the standard normal distribution, C is the confidence level, and set at 0.95.

Table 1.1: Waveform, spectrogram, and SNR of selected six frog species from David Stewart's CD

Frog species	Waveform	Spectrogram	SNR (dB)
Bufo marinus			19.35
Litoria caerulea			15.78
Litoria fallax			43.7
Litoria gracillenta			25.8
Litoria latopalmata			35.85
Litoria rubella			36.2

1.2.4 Acoustic event and background noise

An acoustic event is a localised region of high intensity in a spectrogram. As we can see from Figure. 2.3, there are lots of acoustic events in an one-minute recording. This study focuses

Table 1.2: Waveform, spectrogram, and SNR of eight frog species (recordings from JCU)

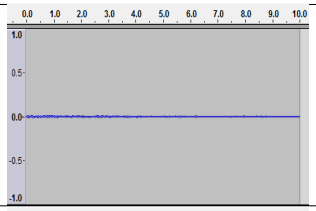
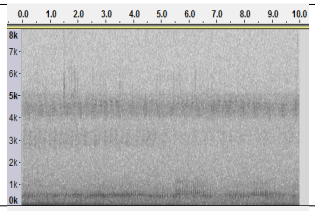
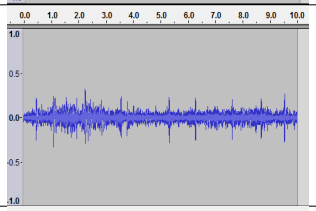
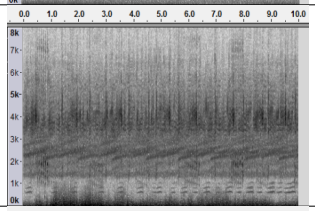
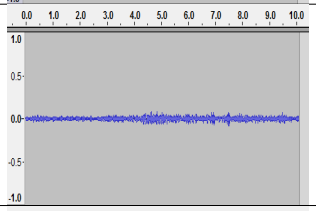
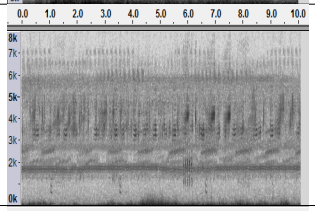
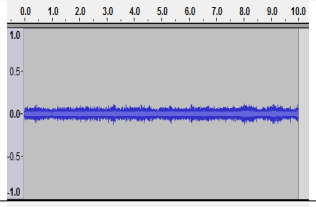
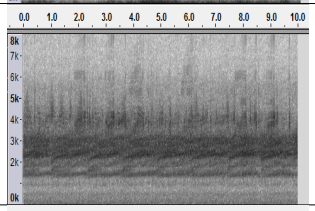
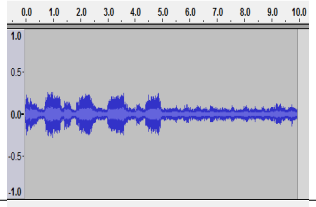
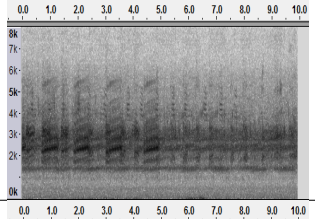
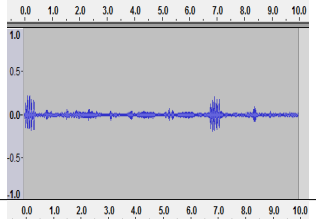
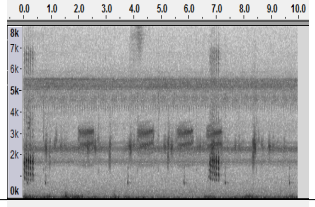
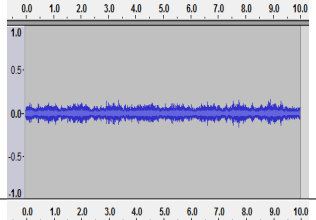
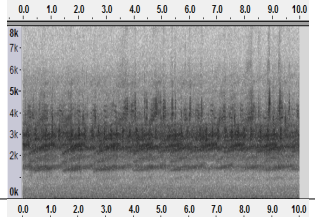
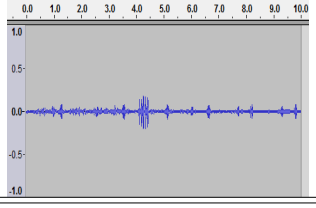
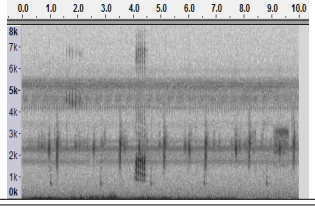
	Waveform	Spectrogram	SNR (dB)
Bufo marinus			1.86
Cyclorana novaehollandiae			-0.13
Limnodynastes terraereginae			-2.88
Litoria fallax			1.52
Litoria nasuta			2.14
Litoria rothii			10.24
Litoria rubella			1.08
Uperolela mimula			10.28

Table 1.3: Averaged parameters of ten syllables of six frog species (David Stewart's CD)

Parameters / Frog species	Syllable duration (milliseconds)	Dominant frequency (Hz)	Oscillation rate (cycle/second)
Bufo marinus	NA	600 ± 30	15 ± 5
Litoria caerulea	500 ± 30	500 ± 75	50 ± 10
Litoria fallax	430 ± 25	4700 ± 450	70 ± 10
Litoria gracillenta	430 ± 25	4700 ± 450	70 ± 10
Litoria latopalmata	30 ± 5	1400 ± 120	5 ± 2
Litoria rubella	160 ± 15	4100 ± 380	70 ± 10

Table 1.4: Confidence interval of signal and noise (David Stewart's CD)

Parameters / Frog species	Confidence intervals of signal	Confidence intervals of noise
Bufo marinus	-9.27*10^{-6} $\pm$ 2.10*10^{-3}	-2.67*10^{-5} $\pm$ 2.26*10^{-4}
Litoria caerulea	-6.73*10^{-5} $\pm$ 2.40*10^{-3}	-6.89*10^{-5} $\pm$ 3.90*10^{-4}
Litoria fallax	-5.85*10^{-6} $\pm$ 1.50*10^{-3}	-2.73*10^{-5} $\pm$ 9.62*10^{-6}
Litoria gracillenta	-7.12*10^{-5} $\pm$ 2.00*10^{-3}	-7.70* 10^{-5} $\pm$ 1.00*10^{-4}
Litoria latopalmata	-6.58*10^{-5} $\pm$ 2.70*10^{-3}	-1.02*10^{-4} $\pm$ 4.36*10^{-5}
Litoria rubella	-3.13*10^{-5} $\pm$ 3.00*10^{-3}	-9.87*10^{-5} $\pm$ 4.70*10^{-5}

Table 1.5: Confidence interval of signal and noise for JCU recordings

Parameters / Frog species	Confidence intervals of signal	Confidence intervals of noise
Bufo marinus	-1.36*10^{-5} $\pm$ 6.00*10^{-5}	-3.22*10^{-5} $\pm$ 4.80*10^{-5}
Cyclorana novaehollandiae	1.30*10^{-3} $\pm$ 8.70*10^{-4}	1.30*10^{-3} $\pm$ 8.90*10^{-4}
Limnodynastes terraereginae	6.43*10^{-5} $\pm$ 1.30*10^{-4}	1.86*10^{-4} $\pm$ 1.89*10^{-4}
Litoria fallax	2.31*10^{-5} $\pm$ 5.00*10^{-4}	5.75*10^{-6} $\pm$ 4.22*10^{-4}
Litoria nasuta	-1.55*10^{-4} $\pm$ 4.90* 10^{-4}	6.25* 10^{-6} $\pm$ 3.85*10^{-4}
Litoria rothii	-3.32*10^{-4} $\pm$ 9.2*10^{-4}	1.61* 10^{-4} $\pm$ 2.84*10^{-4}
Litoria rubella	-8.15*10^{-5} $\pm$ 5.52*10^{-4}	-2.69* 10^{-5} $\pm$ 4.87*10^{-4}
Uperolela mimula	-1.20*10^{-3} $\pm$ 1.10*10^{-3}	-9.36*10^{-4} $\pm$ 3.35*10^{-4}

on the frog vocalisations, and frog calls are recorded as signals. Consequently, all the other events are called background noise. In this study, both high and low SNR recordings are investigated to build a robust frog call classification system. Most previous studies present the frog call classification system using high SNR recordings. The high SNR recordings often

assume that there is only one frog species in each individual recording with few background noises ($SNR \geq 15dB$). In contrast, most low SNR recordings consist of more than one frog species in an individual recording with lots of background noises ($SNR \leq 15dB$). For the low SNR recordings, Table 1.6 show the power spectral density of signal and noise. It can be seen that the noises in low SNR recordings are often generated by several sources and broadband, which cover different frequency bands and lead to the frequency overlapping between the signal and noise. Therefore, it is challenging to improve the classification performance in high SNR recordings over current frog call classification systems, and analyse the recordings containing background noise and simultaneous vocalising events.

1.2.5 Frog call classification

For a frog call classification system, it often consists of four parts (Figure. 1.4) : (1) pre-processing, which includes signal processing and noise reduction; (2) syllable segmentation, which is used to generate basic classification unit for frog calls; (3) feature extraction; (4) classification.

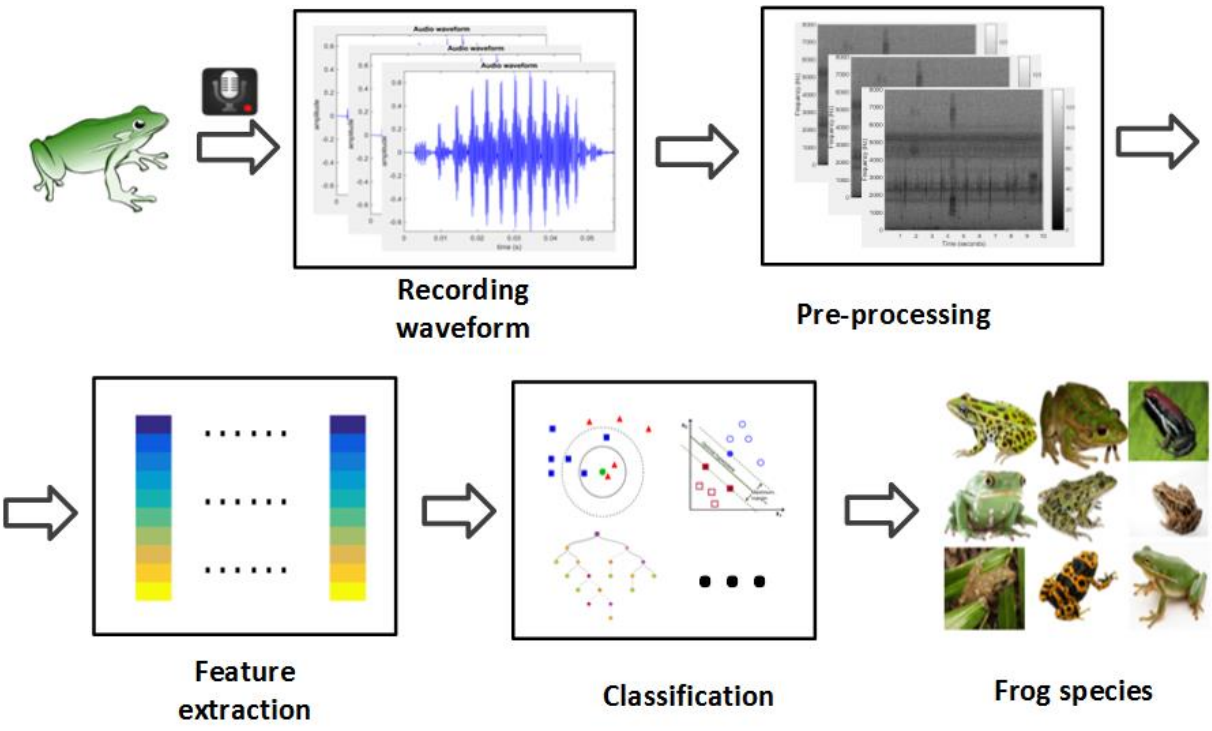
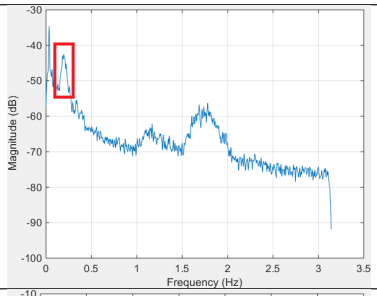
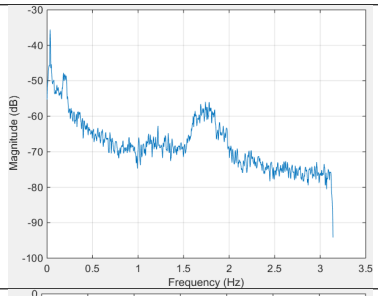
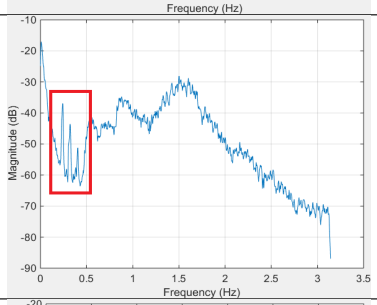
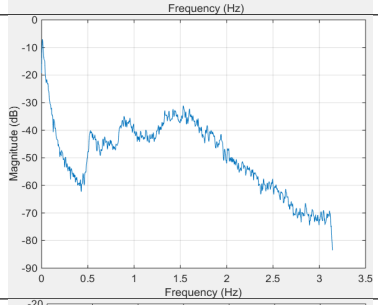
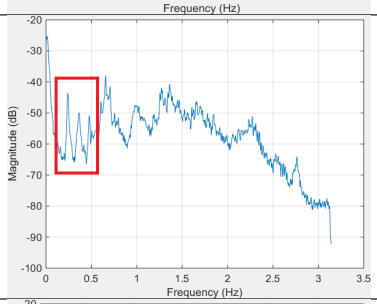
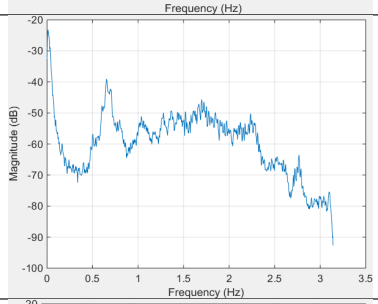
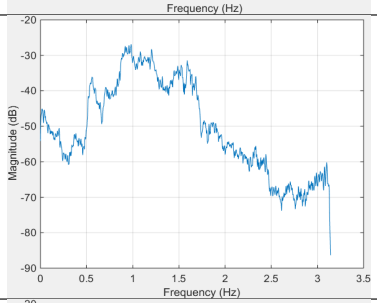
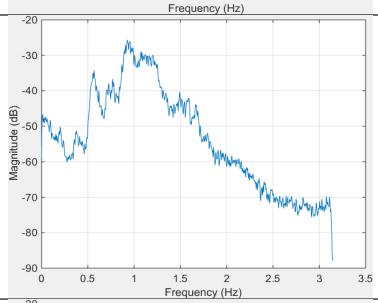
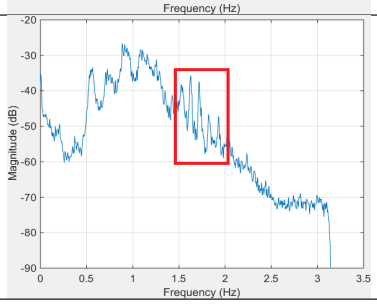
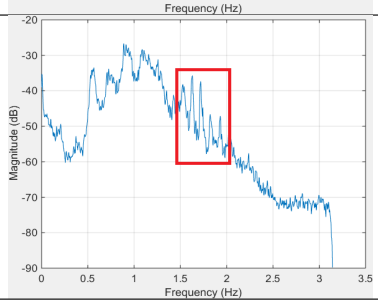
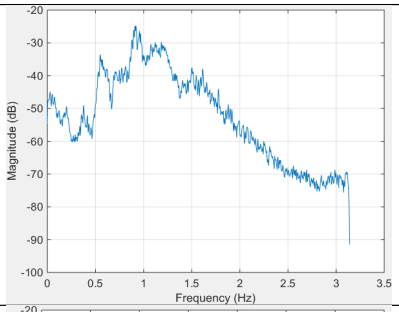
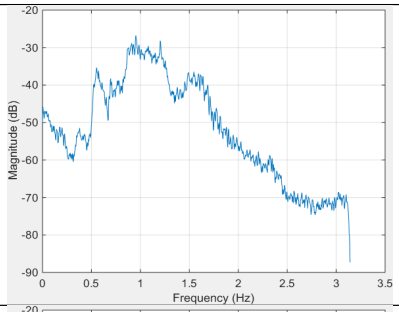
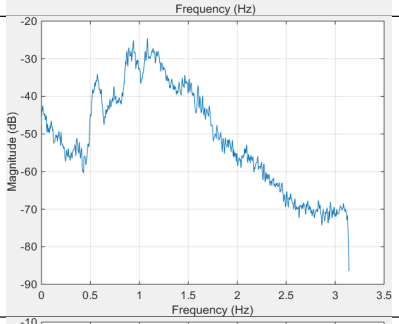
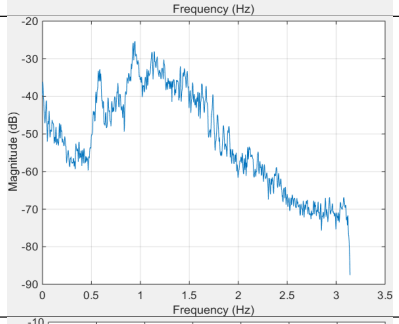
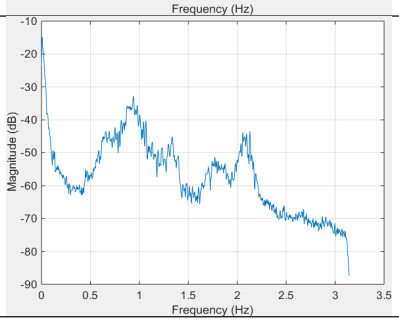
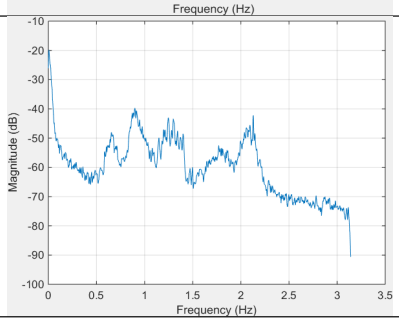


Figure 1.4: Flowchart of frog call classification

Table 1.6: Power spectral density (PSD) estimate of signal and noise (JCU recordings); for some frog species, the PSD difference between the signal and background noise is marked with the red rectangle, which indicates the frequency location of specific frog species; for others, the PSD of signal and noise is very similar, which means that some sources have similar frequency information with frog species

Parameters / Frog species	Welch PSD (Signal)	Welch PSD estimate (Noise)
Bufo marinus		
Cyclorana novaehollandiae		
Limnodynastes terraereginae		
Litoria fallax		
Litoria nasuta		

Litoria rothii		
Litoria rubella		
Uperolela mimula		

1.2.6 Research problem

Most datasets used in previous frog call classification studies assume that there is only one frog species in each individual recordings. However, these resulting frog calls cannot reflect the characteristics of frog vocalisations in real-world situations, such as background noise, frog chorus, and other vocalising animals. To develop a robust frog call classification system for environmental recordings, two main challenges have been identified.

Challenge 1: Most previous work studied frog call classification using high SNR recordings, the first challenge is thus to further improve the classification performance using high SNR recordings. Various acoustic features have been investigated for the classification of frog calls in high SNR recordings. Since our work finally aims to classify frog calls in low SNR recordings, most features that can successfully classify high SNR recordings cannot perform well for low SNR recordings. Consequently, it is still a big challenge to develop robust acoustic features to classify frog species in low SNR recordings.

Challenge 2: Another challenge is the classification framework for studying frog vocalisations

in low SNR recordings. Since most previous work assumed that each individual recording consists of only one frog species, a single-instance single-label (SISL) framework is suitable for classifying frog calls in those high SNR recordings. However, low SNR recordings often have different characteristics containing more than one frog species, the SISL framework is no longer suitable. Therefore, different classification frameworks need to be investigated to study frog vocalisations in low SNR recordings.

1.2.7 Research questions

The research questions are developed in order to solve the aforementioned problems, which can be categorised into three parts.

1. which acoustic features used for addressing high SNR recordings can be transplanted to study low SNR recordings?

2. How to develop acoustic features to classify frog calls in low SNR recordings?

3. How to employ suitable classification frameworks to classify multiple simultaneous vocalising frog species in low SNR recordings?

1.2.8 Aims and objectives

This thesis aims to develop a robust frog call classification system to monitor the environment. For high SNR recordings, we want to improve the classification performance. As for the low SNR recordings, we plan to design novel frameworks to classify multiple simultaneously vocalising frog species. With our classification results, ecologists can then make decisions on how to protect and improve the health of frog populations. The specific research objectives are listed below.

1. To improve the current representation schemes for modelling frog calls in high SNR recordings

2. To develop robust feature extraction methods for frog call classification in low SNR recordings

3. To investigate machine learning techniques (MIML learning and ML learning) to tackle the frog call classification problem in low SNR recordings

1.2.9 Significance and contributions

For the development of sensor techniques, acoustic sensors have been widely deployed in the field for surveying vocalising animals. Different from recordings collected in the constrained environment, recordings collected in the field often have low SNR and consist of multiple simultaneous vocalising frog species. In this dissertation, we first investigate the high SNR recordings to further improve the classification performance of frog recordings with high SNR. Then, those features that can be used for studying frog calls in low SNR recordings are transplanted from high SNR recordings for further analysis. Since field recordings often consists of multiple simultaneous vocalising frog species, both MIML and ML learning are used for the classification of those low SNR field recordings. Meanwhile, the frog calling activity can also be monitored based on the MIML and ML classification results. With our developed frog call classification frameworks, ecologists can then analyse frogs by collecting audio data. It will significantly reduce the expert labour cost for monitoring frog calling activity of a particular area. The monitoring result can also help reveal the importance of environment protection, which can be achieved via studying the correlation between the frog calling activity and weather variables.

1.2.10 Thesis structure

This thesis consists of eight chapters. Chapter 1 described the background, motivation and contributions of the thesis. Chapter 2 reviews the literature related to frog call classification. Chapter 3 discusses a number of feature extraction approaches in detail. Also various classification techniques used in this research will be explained, so one can understand the strategy employed for each classifier. Chapter 4 discusses the implementation of frog call classification by syllable features. Chapter 5 discusses wavelet analysis for frog call classification. Chapter 6 discusses the use of multiple-instance multiple-label framework for frog call classification. Chapter 7 discusses the multiple-label learning for frog call classification. Chapter 8 concludes this research and recommends possible directions in future work.

Chapter 2

Literature Review

This chapter reviews the extant literature on frog call classification. It aims to give a quantitative and detailed analysis of related techniques used in frog call classification. Since previous studies have not used multiple-instance multiple-label (ML) learning or multiple-label (ML) learning for frog call classification, we just focus on the studies of single-instance single-label (MIML) classification of frog calls.

2.1 Introduction

Over the past decade, frog biodiversity has rapidly declined because of frogs' sensitive to habitat loss and degradation, introduced invasive species, and environmental pollution [Dudgeon et al., 2006]. On one hand, frog biodiversity is rapidly declining, and on the other frogs are greatly valuable for the environment. Firstly, frogs are an integral part of the food web, and the decline of their population can result in negative impacts through the whole ecosystem. Secondly, frogs are famous indicator species for environment health. Finally, frogs are very useful in medical research that benefit human [1]. The rapid biodiversity decline and great importance of frogs make it necessary for frog biodiversity monitoring to increase.

To monitor the change of frog biodiversity and optimise the protection policy, many researchers have shown interest in studying frogs. Compared to counting frogs by visual observation, hearing the vocalisations of frogs is much easier. Consequently, frog vocalisations are often used for monitoring frogs. There are two approaches for acoustic frog monitoring. The

[1] http://www.savethefrogs.com/why-frogs

traditional field survey methods require ecologists to physically visit sites to collect acoustic data, which are both time-consuming and costly. In contrast, recent advances in acoustic sensor techniques have greatly extended the spatio-temporal scale for acoustic monitoring of frog biodiversity [Wimmer et al., 2013]. The large volumes of acoustic data collected this way make it essential to develop new automated methods of analysis.

Over the last few years, many researchers have described automated methods for detecting and classifying frog calls [Camacho et al., 2011, Chen et al., 2012, Gingras and Fitch, 2013, Han et al., 2011, Huang et al., 2014, 2009, 2008, Xie et al., 2015b]. However, there is no paper that summarises those methods. In this work, we present a comprehensive survey of frog call classification to provide acoustic signal researchers with basic information, current methods and trends in this field.

Three parts play important roles in the performance and precision of frog call classification: signal pre-processing, feature extraction, and classification. In this survey, these three important parts of frog call classification are presented as shown in Fig. 3.2.

Signal pre-processing consists of signal processing, noise reduction and syllable segmentation. Signal processing often denotes changing a signal from one-dimension (audio data) into two-dimensional representation (image). Noise reduction is essential to improve the classification performance. Since the elementary acoustic unit for frog call classification is the syllable, which is a continuous vocalization emitted by an individual, segmenting continuous recordings of frog calls into individual syllables is necessary.

Previous studies have developed various methods for feature extraction [Camacho et al., 2011, Chen et al., 2012, Gingras and Fitch, 2013, Han et al., 2011, Huang et al., 2014, 2009, 2008, Xie et al., 2015b]. Here we review and analyse all the used features: time domain and frequency domain features, time-frequency features, cepstral features, and other features. After feature extraction, numerous classifiers have been proposed for frog call classification. A summary of those classifiers is given in section 2.4.

It is worth noting that most previous researchers used different databases for their experiments because frog call research is often related to geographical regions [Jang et al., 2011]. Consequently, there is still a lack of uniformity in the way classification methods are evaluated and assessed. This survey is not meant to compare all previous frog call classification methods and find the best one, but to assemble all the methods to provide other researchers with a

direction for the classification of frog calls. To be specific, we mainly survey different features used for frog call classification because most studies focus on new features rather than new signal processing techniques, syllable segmentation methods, or classifiers.

The remainder of this survey is organised as follows: In section 2.2, signal pre-processing is presented in its three parts: signal processing, noise reduction, and syllable segmentation. In section 2.3, different acoustic features are investigated for frog call classification. In section 2.4, numerous classifiers are studied for frog call classification. In section 2.5, experimental results of state-of-the-art research are discussed. Finally, the discussion and conclusion are given in sections 2.6 and 2.7, respectively.

2.2 Signal pre-processing

For frog call classification, signal pre-processing is the first step after acoustic data is collected. It often consists of signal processing, noise reduction, and syllable segmentation. Each part of signal pre-processing is described below.

2.2.1 Signal processing

Signal processing often denotes the transformation of frog calls from one-dimension (recording waveform) to two dimensions (time-frequency representation). Many techniques have been developed for this transformation including short-time Fourier transform (STFT) [Allen, 1977], Wigner-Ville distribution [Boashash and Black, 1987], and wavelet transform [Meyer and Salinger, 1995]. STFT is the most widely used technique among them for its flexible implementation and better applicability. Given one frog call $x(t)$, its fast Fourier transform can be expressed as

$$X(k) = \sum_{n=0}^{L-1} x(n)w(n)e^{-j2\pi kn/L}, 0 \leq k \leq L-1 \tag{2.1}$$

where $X(k)$ is the frequency domain signal (spectrum) and denotes each frame of the spectrogram, and $w(n)$ is the window function. The waveform, spectrum and spectrogram of one individual syllable for *Mixophyes fasciolatus* is illustrated in Fig. 2.1. Here three representations are consistent with features in three domains: the time domain, frequency domain and time-frequency domain.

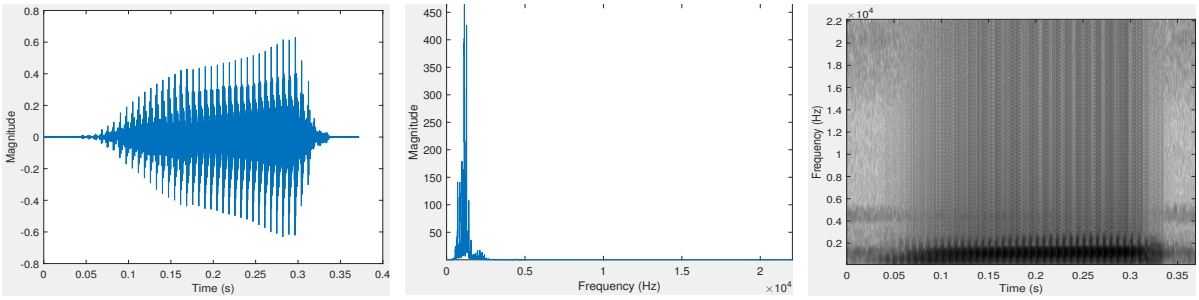


Figure 2.1: Waveform, spectrum and spectrogram of one frog syllable for *Mixophyes fasciolatus*. The window function, size and overlap are Hamming window, 128 samples and 85%, respectively

2.2.2 Noise reduction

Noise reduction is an optional process for frog call classification. Huang et al. [2014] applied a de-noise filter for noise reduction. The wavelet threshold function in the one-dimensional signal was used as the filter kernel function. Bedoya et al. [2014] introduced spectral noise gating method for noise reduction. Specifically, the selected frequency band spectrum of the frogs' call to be detected was estimated and suppressed. Xie et al. [2015d] used Wiener filtering to remove the background graininess, then applied spectral subtraction to the filtered spectrogram using a modified method from the adaptive level equalization algorithm. While the aforementioned noise reduction methods can remove some background noise, some of the desired signals will be suppressed. Noise reduction methods are therefore selectively used based on signal-to-noise ratio of the acoustic data and the research problem.

2.2.3 Syllable segmentation

For frog calls, the basic elementary acoustic unit is a syllable, which is a continuous frog vocalisation emitted by an individual frog [Huang et al., 2009]. The precision of syllable segmentation will directly affect the classification performance, since features used for frog call classification are calculated based on each syllable. Frog syllable segmentation methods in previous studies are summarised and listed in Table 2.1. All previous methods except [Xie et al., 2015c] cannot address recordings with simultaneous vocalising frog calls. Meanwhile, those methods that used the time domain feature for segmentation, cannot address recordings with low signal-to-noise ratio. Xie et al. [2015c] introduced an unsupervised learning method, which included multiple image processing techniques for syllable segmentation. However, the

downside of this unsupervised process was that not all segmented syllables correspond to frog vocalisations.

Table 2.1: Summary of related work for frog syllable segmentation. Here, E denotes energy, ZCR denotes zero-crossing rate.

Authors	Features for segmentation	Procedure
Han et al. [2011]	Spectral entropy	Manual
Jaafar and Ramli [2013]	E and ZCR	Sequential
Huang et al. [2009]	Amplitude	Non-sequential
Chen et al. [2012]	Spectrogram	Non-sequential
Xie et al. [2015b]	Spectrogram	Non-sequential
Colonna et al. [2015]	Incremental E and Incremental ZCR	Sequential and real time
Xie et al. [2015c]	Image processing	Non-sequential

2.3 Acoustic features for frog call classification

Developing effective acoustic features that show greater variation between rather than within species is important for achieving robust classification results [Fox, 2008]. For frog call classification, acoustic features can be classified into four categories: time domain and frequency domain features, time-frequency domain features, cepstral features, and other features.

2.3.1 Time domain and frequency domain features for frog call classification

Time domain features for frog call classification have been explored for a long time [Camacho et al., 2011, Chen et al., 2012, Dayou et al., 2011, Huang et al., 2014, 2009, 2008]. Time domain features are often combined with frequency domain features for frog call classification.

Huang et al. [2009] used spectral centroid, signal bandwidth, and threshold-crossing rate for frog call classification with a k-nearest neighbour classifier (K-NN) and support vector machines (SVM). In another work, Huang et al. [2014] combined spectral centroid, signal bandwidth, spectral roll-off, threshold-crossing rate, spectral flatness, and average energy to classify frog calls using neural networks. Another paper published by [Huang et al., 2008] used spectral centroid, signal bandwidth, spectral roll-off, and threshold-crossing rate for frog call classification. Dayou et al. [2011] combined Shannon entropy, Rényi entropy and Tsallis

entropy for frog call classification. Based on this work, Han et al. [2011] improved the classification accuracy by replacing Tsallis entropy with spectral centroid. To classify anurans into four genera, a three-parameter model was proposed based on advertisement calls, [1] which used mean values for dominant frequency, coefficients of variation of root-mean square energy, and spectral flux [Gingras and Fitch, 2013]. With this model, three classifiers were employed for classification: K-NN, a multivariate Gaussian distribution model and a Gaussian Mixture Model (GMM) [Gingras and Fitch, 2013]. Chen et al. [2012] proposed a method based on syllable duration and a multi-stage average spectrum for frog call recognition. Their recognition stage was completed by the Euclidean distance-based similarity measure. Camacho et al. [2011] used the loudness, timbre and pitch to detect frogs with a multivariate ANOVA test.

2.3.2 Time-frequency features for frog call classification

For frog call classification, we often transform the one-dimensional signal into its two-dimensional time-frequency representation. Then features based on the time-frequency representation can be calculated for classification. Acevedo et al. [2009] developed two feature sets for automated animal classification. The first was minimum and maximum frequencies, call duration, and maximum power; the second was minimum and maximum frequencies, call duration, and frequency of maximum power in eight segments of duration. With two feature sets, three classifiers were used for the classification: linear discriminant analysis(LDA), decision tree and SVM. Brandes [2008] proposed a method for classifying animals using duration, maximum frequency, and frequency bandwidth, and with Hidden Markov Model (HMM) used as the classifier. Yen and Fu [2002] combined wavelet packet feature extraction and two different dimensionality reduction algorithms to produce the final feature vectors. Then, they adopted a neural network classier for classification. Grigg et al. [1996] developed a system to monitor the effect on frog population of Queensland of the introduced Cane Toad. The classification was based on the local peaks in the spectrogram using Quinlan's machine learning system, C4.5. Brandes et al. [2006] proposed a method to classify frogs using central frequency, duration, and bandwidth with a Bayesian classifier. Croker and Kottege [2012] introduced a feature vector for detecting frogs with a similarity measure based on Euclidean distance. The feature vector consisted of dominant frequency, frequency difference between the lowest and dominant

[1] an advertisement call is produced by a male frog in order to attract females during the breeding season and to warn other rival males of his presence.

frequencies, frequency difference between the highest and dominant frequencies, time from the start of the sound to the peak volume, and time from the peak volume to the end of the sound. Xie et al. [2015b] developed a method for frog call classification using syllable duration, dominant frequency, oscillation rate [2], frequency modulation, and energy modulation using a K-NN classifier.

2.3.3 Cepstral features for frog call classification

Cepstral features are also popular for frog call classification. These features include Linear Prediction Coefficients (LPCs), Mel-frequency cepstral coefficients (MFCCs), and perceptual wavelet packet decomposition sub-band cepstral coefficients (PWSCCs) [Xie et al., 2015a]. Colombia and del Cauca [2009] introduced LPCs for frog call classification with a modified K-Means classifier. Jaafar et al. [2013a] introduced MFCCs and LPCs as features, and K-NN and SVM as classifiers for frog call identification. Yuan and Ramli [2012] also used MFCCs and LPCs as features, and K-NN as the classifier for frog sound identification. Lee et al. [2006] used the averaged MFCCs and LDA for the automatic recognition of animal sounds. Bedoya et al. [2014] combined MFCCs and a learning algorithm for multivariate data analysis (LAMDA) for frog call recognition. Vaca-Castano and Rodriguez [2010] proposed a method to identify animal species, which consisted of MFCCs, principal component analysis (PCA) and K-NN. Jaafar and Ramli [2013], Jaafar et al. [2013b], Tan et al. [2014] published three papers about frog call classification with MFCCs, Δ MFCC and $\Delta\Delta$ MFCC calculated as features, and K-NN and SVM used as classifiers. Colonna et al. [2012] introduced MFCCs for classifying anurans with a K-NN classifier. Xie et al. [2015a] proposed a novel feature set named perceptual wavelet packet decomposition sub-band cepstral coefficients for frog call classification. Compared with MFCCs, this feature set was more suitable for the frequency distribution of frog calls and provided a better performance for classifying frog calls . Noda et al. [2016] fused time domain features with cepstral features for frog call classification which achieved a better classification performance than using only cepstral features. Three classifiers were investigated for the classification: HMM, random forest, and SVM.

[2] Oscillation rate denotes the number of pulses within one second.

2.3.4 Other features for frog call classification

Besides time domain features, frequency domain features, time-frequency domain features and cepstral features, other features are also introduced to classify frog calls. Wei et al. [2012] proposed a distributed sparse approximation method based on $\ell 1$ minimization for frog call classification. Dang et al. [2008] extracted vocalization waveform envelope as features, then classified calls by matching the extracted envelope with the original signal envelope. Xie et al. [2015d] used two feature sets for frog call classification: (1) minimum frequency, maximum frequency, bandwidth, duration, acoustic event area, acoustic event perimeter, acoustic event non-compactness, acoustic event rectangularity. (2) frequency mean, frequency variance, frequency skewness, frequency kurtosis, time mean, time variance, time skewness, time kurtosis, mask mean, mask standard deviation. Feature set (1) was used to describe the mask of each segmented event, feature set (2) was used to describe the statistical properties of each segmented event, each event corresponds to an individual event in Fig. 2.2. Meanwhile, Xie et al. [2015c] introduced ridge related features for frog call classification: mean value for dominant frequency, low and high frequencies, histogram of ridges, and entropy of ridges in horizontal and vertical directions.

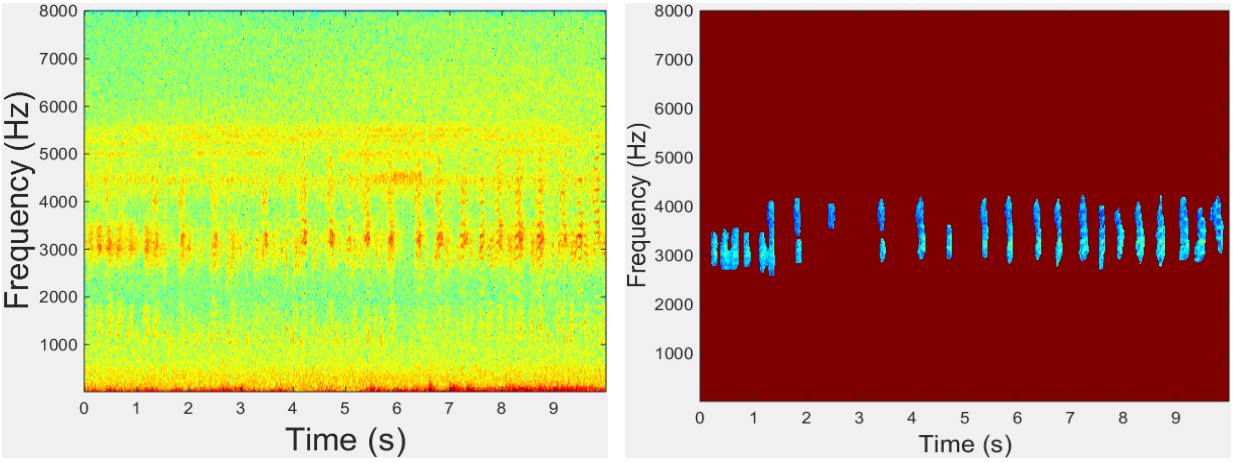


Figure 2.2: Original spectrogram and segmented events after applying acoustic event detection to the original spectrogram image.

2.4 Classifiers

For frog call classification, numerous pattern recognition methods have been used to construct the classifier, such as the Bayesian classifier [Brandes et al., 2006], k-nearest neighbour classifier (K-NN) [Colonna et al., 2012, Dayou et al., 2011, Gingras and Fitch, 2013, Han et al., 2011, Huang et al., 2009, 2008, Jaafar and Ramli, 2013, Jaafar et al., 2013a,b, Vaca-Castano and Rodriguez, 2010, Xie et al., 2015a,b,d, Yuan and Ramli, 2012], support vector machine (SVM) [Acevedo et al., 2009, Gingras and Fitch, 2013, Huang et al., 2009, 2008, Jaafar et al., 2013a, Tan et al., 2014, Xie et al., 2015c], hidden Markov model (HMM) [Brandes, 2008], Gaussian mixture model (GMM) [Gingras and Fitch, 2013, Huang et al., 2008], neural networks (NN) [Huang et al., 2014, Yen and Fu, 2002], decision tree (DT) [Acevedo et al., 2009, Grigg et al., 1996], one-way multivariate ANOVA [Camacho et al., 2011], and linear discriminant analysis (LDA) [Acevedo et al., 2009, Lee et al., 2006]. Besides classifiers, other methods for classifying frog species include those based on the similarity measure [Chen et al., 2012, Croker and Kottege, 2012, Dang et al., 2008] and those based on the clustering technique [Bedoya et al., 2014, Colombia and del Cauca, 2009, Wei et al., 2012]. K-NN is the most commonly used classifier for its simplicity and easy application. However, the K-NN classifier is sensitive to the local structure of the data, as well as to the initial cluster centroids. Therefore, the K-NN classifier is often run multiple times based on different initial points. SVM is another classifier which is widely used for its good generalization ability. However, the performance of SVM can be quite sensitive to the selection of the regularization and kernel parameters, and it is possible to over-fit when tuning these hyper-parameters. Therefore, selecting suitable parameters for SVM is very important and is realized by grid search in most previous studies [Hsu et al., 2003].

2.5 Experiment results of the state-of-the-art methods

2.5.1 Evaluation criteria

Accuracy is the most widely used statistical criterion for evaluating frog call classification. Other evaluation criteria such as precision, recall, sensitivity, specificity, F-measure, and ROC curves are also used. Before defining these evaluation criteria, we first define true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP) as described by [Gordon et al.,

2003] (1) TP: correctly recognized positives; (2) TN: correctly recognized negatives; (3) FN: positives recognized as negatives; (4) FP: negatives recognized as positives. Then, accuracy, precision, recall (sensitivity), and specificity can be defined as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

$$Specificity = \frac{TN}{FP + TN} \tag{2.5}$$

2.5.2 Experiment results for summarise

Table 2.2 shows the list of summarised frog call classification methods, together with the database they used and corresponding performance.

2.6 Discussion and future work

In this section, each part of a frog call classification system is discussed to give a direction for future work.

2.6.1 Database

One major problem for frog call classification is the lack of an universal database. The databases used are often related to geographical regions, since researchers from different countries focus on particular frog species in their specific area (Table 2.2). Therefore, it is difficult for researchers to compare their particular classification methods. Current studies often focus on the study of limited number of frog species (less than 100), but the number of known amphibian species is above 7000. To reach a high quality resolution, there still is a long way to go.

Table 2.2: A brief overview of frog call classification performance. The asterisk denotes that frog species are not the only animal species to be studied.

Database	Performance	Reference	Data source
3 frog species with 635 calls	Precision of 99%, recall of 92%	Camacho et al. [2011]	Collected from Costa Rica (unavailable)
1 frog species with 100 samples	Sensitivity of 0.85 with specificity of 0.92 when distinguishing *Mixophyes iteratus* calls from other species' call. Sensitivity of 0.88 with specificity of 0.82 against background noise	Croker and Kottege [2012]	Recorded next to a running stream (unavailable)
17 animal types	50% true positive accuracy, over 50 false-negative for 4 animal types	Brandes et al. [2006]	Collected from NE Costa Rica (unavailable)
22 frog species	NA	Grigg et al. [1996]	Collected from Queensland, Australia (unavailable)
4 frog species with 66 samples	Best performance with averaged classification Accuracy of 72.18% and 0.76% for standard deviation.	Yen and Fu [2002]	Unknown
10 frog species, 9 bird species, and 8 cricket species	Accuracy of 88% for frogs	Brandes [2008]*	Collected from NE Costa Rica (unavailable)
9 frog species and 3 bird species with 10061 samples	Best true positive rate of 94.95% and 0.94% for false positive rate	Acevedo et al. [2009]*	Collected from 14 montane sites in Puerto Rico
9 frog species with 90 syllables	Averaged classification accuracy of 90.00%	Dayou et al. [2011]	Obtained from http://www.Frogsaustralia. net.au/frogs
9 frog species with 54 syllables	Averaged classification accuracy of 98.00%	Han et al. [2011]	Obtained from http://www.Frogsaustralia. net.au/frogs
5 frog species with 727 syllables	Averaged classification accuracy of 95.86%	Huang et al. [2008]	Unknown
142 species belonging to four genera	Genus classification accuracy above 70%	Gingras and Fitch [2013]	obtained from commercially available compact disks (CDs) (available)
18 frog species with 960 syllables	Classification accuracy of 94.3%	Chen et al. [2012]	Recorded in a wild field located in the Shan-Ping forest ecological garden in Kaohsiung city, Taiwan (unavailable)
13 frog species with 1514 samples	Averaged recognition rate of 93.4%	Huang et al. [2014]	Unknown
5 frog species with 959 samples	Averaged classification accuracy of 90.03%	Huang et al. [2009]	Unknown
15 frog species with 286 samples	Averaged classification accuracy of 95.67%	Tan et al. [2014]	recorded at Sungai Sedim, in Kulim, Kedah, Malaysia
8 frog species with 160 samples	averaged classification accuracy of 98.1%	Yuan and Ramli [2012]	Obtained from AmphibiaWeb (http://amphibiaweb.org/)(available)
10 frog species with 250 syllables	Averaged classification accuracy of 98.8%	Jaafar et al. [2013a]	Internet database (http://learning.frogbome.org/) and IBM,USM (http://www.frogwatch.org.au/?action=animal.list) (available)
15 frog species with 386 syllables	Averaged classification accuracy of 85.78%	Jaafar et al. [2013b]	Recorded from locations around Baling and Kulim, Kedah, Malaysia (unavailable)
12 frog species with 291 syllables	Averaged classification accuracy of 97%	Jaafar and Ramli [2013]	Recorded from locations around Baling and Kulim, Kedah, Malaysia (unavailable)
12 frog species with 379 samples, 10 bird species with 193 samples	Averaged classification accuracy of 86.6%	Vàca-Castano and Rodriguez [2010]	Recorded in Puerto Rico (http://www.amazon.com/Los -Antibios-Reptiles-Puerto-Rico/dp/0847702 43X) (available)
13 frog species with 916 calls	Averaged classification accuracy of 100%, and 99.61% respectively for two database	Bedoya et al. [2014]	Provided by the Smithsonian Tropical Research Institute (STRI) and the Grupo Herpetolgico de Antioquia (GHA) (unavailable)
30 frog species and 19 cricket calls	Averaged classification accuracy of 96.8% and 98.1%	Lee et al. [2006]	Derived from compact disk (unavailable)
15 frog species with 896 syllables	Precision of 99.00%	Colonna et al. [2015]	Obtained from Internet(http:// bit.ly/1b8bvyE) (available)
10 frog species with 516 syllables	Averaged classification accuracy of 97.45%	Xie et al. [2015a]	Collected from compact disk (http://www.naturesound.com.au/) (available)
15 frog species with 436 syllables	Averaged classification accuracy of 74.73%	Xie et al. [2015c]	Collected from compact disk (http://www.naturesound.com.au/) (available)
16 frog species with 898 syllables	Averaged classification accuracy of 90.5%	Xie et al. [2015b]	Collected from compact disk (http://www.naturesound.com.au/) (available)
14 frog species with 985 syllables	Averaged classification accuracy of 87.00%	Xie et al. [2015d]	Collected from compact disk (http://www.naturesound.com.au/) (available) and one public website (http://amphibiaweb.org/maps/index.html)
9 frog species with 49 samples	Averaged classification accuracy of 97.60%	Colonna et al. [2012]	Collected on the campus of the Federal University of Amazonas in Manaus, Brazil (unavailable)
3 frog species with 50 samples	Averaged classification accuracy of 90%	Dang et al. [2008]	Unknown
1564 syllables of 41 anurans, 5201 syllables of 58 frogs, 10905 syllables of 100 anurans, and 17671 syllables of 199 anurans	98.8%, 96.9%, 95.48%, and 95.38% respectively	Noda et al. [2016]	AmphibiaWeb(41 anurans), 58 frogs from Cuba, 100 anurans from Brizil-Uruguay, and 199 anurans from all datasets (http://www.nhbs.com/) (available)
18 frog species from commercial recordings and field recordings of 8 frog speciesfrom James Cook University recordings	99.5% and 97.4% for 18 frog species and 8 frog species respectively	Xie et al. [2016]	David Stewart's commercial CD (http://www.naturesound.com.au/) and frog calls collected from the wild (https://www.ecosounds.org/) (available)

2.6.2 Signal pre-processing

Currently, short-time Fourier transform (STFT) is the most widely used technique for frog call classification. However, STFT has a trade-off between time and frequency resolution, which restricts the discriminability of features extracted from the spectrogram. In contrast, wavelet packet decomposition (WPD) has a better frequency domain resolution than STFT. The main disadvantage of WPD is the time dependence.

Noise reduction is an optional processing step in frog call classification. For some databases of studies shown in Table 2.2, frog calls have a high signal-to-noise ratio (SNR), where noise reduction is unnecessary. However, when studying recordings of low SNR, noise reduction is essential for improving the classification performance [Bedoya et al., 2014, Huang et al., 2014]. After noise reduction, both the accuracy of syllable segmentation and feature extraction can be relatively improved.

Frog syllable segmentation based on energy and zero-crossing rate cannot address recordings with low SNR. Meanwhile, this method cannot segment recordings with overlapping frog calls. Recent use of unsupervised learning algorithms opens a path for segmenting overlapping frog syllables with image processing techniques. However, like other unsupervised algorithms, this method has the disadvantage that not all segmented syllables are frog vocalizations [Potamitis, 2015]. Briggs et al used a supervised learning algorithm (Random Forest) for bird call segmentation. However, this method required lots of tagged acoustic data to train the classifier [Tjahja et al., 2015].

For syllable segmentation, time domain features are more sensitive to background noise than frequency domain features, because different frequency components can be separated by transforming the signal from time domain to frequency domain. But time domain features cannot segment those overlapping frog syllables, since time domain features have no ability to separate different frequency components. Compared to time domain features, the use of amplitude-frequency information provides a robust method to segment low SNR recordings. To address those overlapping frog syllables, image processing techniques can be a possible solution.

2.6.3 Acoustic features

Most previous studies directly transplant features developed for speech recognition to analyze frog calls, which might not be suitable. For example, MFCCs are designed for studying speech, which are based on the calculation of a non-linear Mel-scale. However, the Mel-scale is designed for the perceptual scale of pitches judged by listeners rather than frogs. The direct use of speech features will therefore restrict classification performance. Recently, Xie et al. [2015a] used an adaptive frequency scaled wavelet packet decomposition to classify frog calls, and it achieved a better performance than Mel-scaled wavelet packet decomposition. Here, an

adaptive frequency scale was generated by applying K-Means clustering to dominant frequencies, which was more accurate and efficient than using Mel-scale [Xie et al., 2015a]. Most frequency domain features are calculated by directly calculating the statistics over frames, which leads to the loss of temporal information. To add the temporal information of the feature set, time domain features can be combined with frequency domain features to achieve higher classification accuracy. Transforming audio data into its a two dimensional representation (such as a spectrogram) for quick visual analysis, has led to increasing attention being given to image processing techniques for automatically analysing animal calls. Ridges extracted from spectrogram images ware applied to perform frog call classification [Xie et al., 2015c]. Besides ridges, other image features are worth being investigated for frog call classification.

2.6.4 Classifiers

Almost all previous studies assume that each recording has only one frog species, then a single-instance single-label classification framework is adopted to classify frog calls. However, recent advances in acoustic sensor techniques have collected large volumes of acoustic data that have multiple simultaneously vocalising frog species, because different frog species tend to call together to make frog chorus (Figure. 2.3). Based on this characteristic of frog call recordings, the classification problem can be naturally framed as a multiple-instance multiple-label classification or a multiple-label classification problem rather than a single-instance single-label classification.

2.7 Conclusions

The main objective of this survey is to provide a research direction for analysing acoustic signals, especially frog calls. With the use of signal processing and machine learning techniques, different frog species can be classified based on their vocalizations. To achieve this goal, three main parts of a frog call classification system are explained: signal pre-processing, feature extraction, and classification. For each part, current techniques used by different researchers are explored. For signal pre-processing, signal processing, noise reduction, and syllable segmentation are studied respectively. For feature extraction, acoustic features in different domains

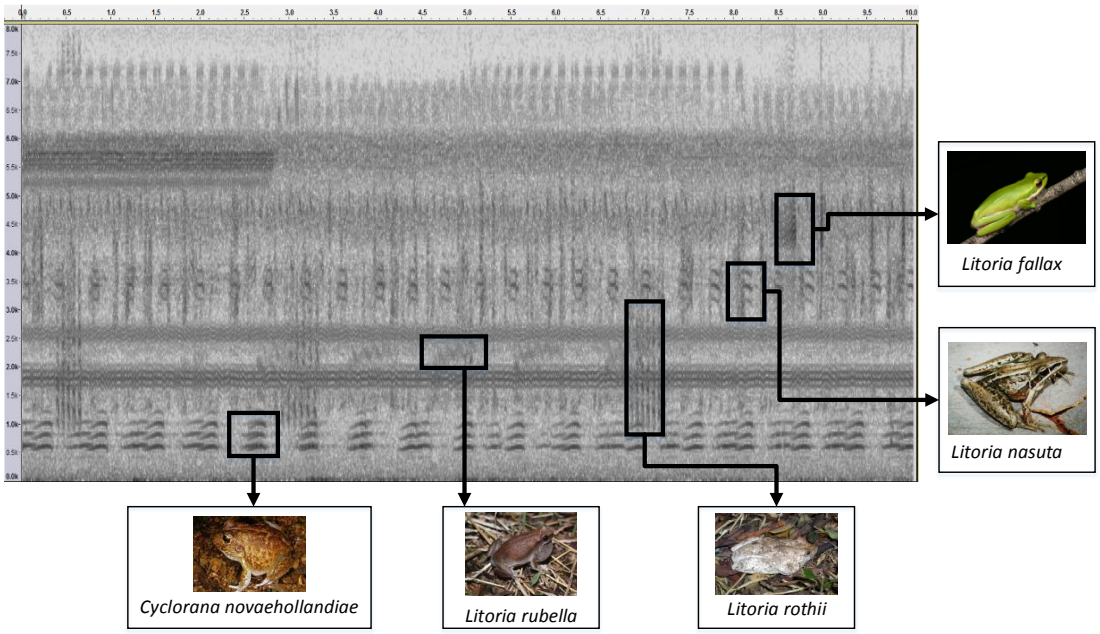


Figure 2.3: An example of collected recording with multiple simultaneously vocalizing frog species. Five frog species exist in this 10-second recording: *Cyclorana novaehollandiae*, *Litoria rubella*, *Litoria nasuta*, *Litoria rothii*, and *Litoria fallax*.

are explored. For classification, different classification frameworks are investigated: single-instance single-label classification, multi-label classification, and multi-instance multi-label classification.

In general, frog call classification is still in its infancy as a field of study, and potential applications and unsolved problems are extending every day. For future work, it is worth further improving the accuracy and efficiency of noise reduction and syllable segmentation because they are critical processes for frog call classification. Since collected frog calls in the field often contain many background noises (birds, insects, rain, wind, human voices, etc.), it is necessary to design new noise reduction methods based on different environments. It is also necessary to develop accurate and efficient methods for syllable segmentation for its great influence in the frog call classification system performance. Currently, studies have focused on frequency domain features for classification. In the future, time domain features can be more incorporated for increasing the accuracy of frog call classification. For the classification frameworks, using MIML learning or ML learning for studying environmental recordings may be a productive research direction because of the characteristic of collected acoustic data. It is also worth making a uniform dataset that covers different frog species from different areas,

since there is still no available uniform datasets of frog calls. Then researchers can evaluate their particular methods on a uniform platform.

Chapter 3

Methodology

The flowchart of a frog call classification system is depicted in Figure. 3.2, which includes three parts: signal pre-processing, feature extraction, and classification. In this dissertation, we primarily focus on feature extraction and classification.

3.1 Feature extraction

In this section, feature extraction which always plays a key role in the classification performance is studied. A number of methods that have been used in this research for extracting relevant features of frog vocalizations are investigated.

3.1.1 Introduction

For any pattern recognition or statistical analysis, feature extraction aims to provide the most compact and informative signature for a machine learning model or classifier. Meanwhile, feature extraction is often seen as a step to facilitate the subsequent learning and generalization parts. For a classification system, it is necessary to perform feature extraction, because analysing complex data generally requires a large amount of memory and computational power. Also, a classification system without feature extraction might be caused to be over-fitting for training samples and generalize poorly to new samples. To be specific, feature extraction consists of a number of steps as shown in Figure. 3.1. To analyse different types of objects, the representation of raw data varies a lot. For instance, an audio and a video signal can be displayed using 1-Dimension representation and 2-Dimension representation, respectively. Since the raw

data are usually collected under different unconstrained environments, it is necessary to perform pre-processing to obtain efficient and distinguished features.

Figure 3.1: Feature extraction process

After pre-processing, feature generation is the next step, which directly applies various standard methods to the data. To increase the discriminability of generated features, it is also worthwhile considering the specific domain knowledge and underlying physical phenomenon. Then, feature set is constructed with the format of a scalar or vector per feature. Also the format can be one vector that concatenates all features or one matrix holding all samples of features. Finally, dimension reduction is an optional step with the following advantages: (1) reduces the time and storage space; (2) removes the multi-collinearity and improves the system performance; (3) makes it easier to visualise the data. As for the implementation of dimension reduction, there are two approaches: (1) selects a subset of the original features; (2) transforms the data into another space using the different dimensions.

The vocalisations of frogs are innate in structure and may therefore contain indicators of phylogenetic history. Thus, frogs that are closely related phylogenetically often share similar advertisement calls. Based on this assumption, frogs can be classified by their calls. Over the past two decades, different semi-automated and automated methods have been developed to extract features from frog calls. Most features that have been used in the previous studies are list below.

3.1.2 Various features used in the literature

Spectral centroid

Spectral centroid (SC) 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 magnitude as the weight, it is calculated as the weighted mean of the frequencies.

$$SC = \frac{\sum_{k=0}^{N-1} f_k X(k)}{\sum_{k=0}^{N-1} X(k)} \tag{3.1}$$

where $X(k)$ is the DFT of the signal syllable of the k-th sample, N is the half size of DFT.

Spectral flatness

Spectral flatness (SF) provides a way to quantify the tonality of a sound. A high spectral flatness indicated a similar amount of power of the spectrum in all spectral bands. Spectral flatness is measured by the ratio of the geometric mean and the arithmetic mean of the power spectrum and defined as

$$SF = \frac{\sqrt{\frac{1}{N}\sum_{k=0}^{N-1} InX(k)}}{\frac{1}{N}\sum_{k=0}^{N-1} X(k)} \tag{3.2}$$

Spectral flux

Spectral flux is used to measure how quickly the power spectrum of a signal is changing. By comparing the power spectrum between one frame and its previous one, the spectral flux can be obtained. The calculation of spectral flux is denoted as

$$SX = \sum_{k=-N/2}^{N/2-1} H[|X(n,k)| - |X(n-1,k)|] \tag{3.3}$$

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

Spectral roll-off

Spectral roll-off (SR) is often used to measure the spectral shape, and defined as the frequency H below which of the magnitude distribution is concentrated.

$$\sum_{k=1}^{H} X(k) = \theta \sum_{k=1}^{N-1} X(k) \tag{3.4}$$

Here θ is set at 0.85.

Signal bandwidth

Signal bandwidth (BW) is often used to represent the difference between the upper and lower cutoff frequencies.

$$BW = \sqrt{\frac{\sum_{k=0}^{N-1}(k - SC)^2|x(n)|}{\sum_{k=0}^{N-1} X(k)}} \tag{3.5}$$

Mean values for dominant frequency

Dominant frequency (spectral peak) corresponds to the point of maximal amplitude along the frequency spectrum. Mean values for dominant frequency is defined as

$$MDF = \frac{\sum_{i=1}^{K} f_i}{K} \tag{3.6}$$

where K is the number of frames in a frog syllable.

Oscillation rate

Oscillation rate is calculated in the frequency boundary around the fundamental frequency. First, the power within the frequency boundary is calculated. Then, the first and last 20% part of the power vector is discarded after normalization for their uncertain. Next, the autocorrelation is applied by the length of the vector. Furthermore, a discrete cosine transform is employed to the vector after the mean subtraction, and the position of the highest frequency is achieved for calculating the oscillation rate. Detailed description can be found in our previous study [Xie et al., 2015b].

Zero-crossing rate

Zero-crossing rate (ZCR) means the rate of signal change along a signal. When adjacent signals have different signs, a zero-crossing occurs. It can be defined as

$$ZCR = \frac{1}{2}\sum_{k=0}^{N-1}[sgn(X(k)) - sgn(X(k+1))] \tag{3.7}$$

Shannon entropy

Shannon entropy (SEY) is the expected information content of a sequence of signal. It describes the average of all the information contents weighted by their probabilities p_i.

$$SEY = -\sum_{i=1}^{L} p_i log_2(p_i) \tag{3.8}$$

where L is the length of a frog syllable.

Rényi entropy

Rényi entropy can be used to obtain different averaging of probabilities via the parameter $alpha$, and defined as

$$REY = \frac{1}{1-\alpha} log_2(\sum_{i}^{n} p_i^{\alpha}) \tag{3.9}$$

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

Average energy

Average energy (AVG) is defined as the sum of intensity of signal.

$$AVG = \frac{1}{f} \sum_{k=0}^{f-1} X(k)^2 \tag{3.10}$$

LPC

Linear prediction coding (LPC) is often used to represent the spectral envelope of speech sound [Itakura, 1975]. LPC coefficients can be calculated using a linear predictive filter.

$$X(n) = \sum_{i}^{p} a_i x(n-i) \tag{3.11}$$

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).

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 **??**. The steps for MFCCs implementation are list as follows.

Step 1: Pre-emphasis

$$y(n) = s(n) - \alpha s(n-1) \tag{3.12}$$

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.

$$x(n) = w(n) * y(n) \tag{3.13}$$

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

Step 3: Spectral analysis Compute the discrete Fourier transform (DFT) of each frame of the signal. By considering $\omega = 2\pi k/N$, the DFT of each frame of the signal is

$$X(k) = \sum_{n=0}^{N-1} x(n) e_{-j\omega} \tag{3.14}$$

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.

$$E_j = \sum_{k=0}^{N/2-1} \phi_j(k) A_k, 0 \leq j \leq J-1 \tag{3.15}$$

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

$$A_k = |X[k]|^2, 0 \leq k \leq N/2 \tag{3.16}$$

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

E_j.

$$C_m^j = \sum_{j=0}^{J-1} \cos(m\frac{\pi}{J}(j+0.5))log_{10}(E_j), 0 \leq m \leq L-1 \tag{3.17}$$

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 $16(L = 16)$. After calculating MFCCs from each frame, the averaged MFCCs of all frames within one syllable are calculated.

$$f_m = \frac{\sum_{i=1}^{K}(C_m^l)}{K}, 0 \leq m \leq L-1 \tag{3.18}$$

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.

$$\hat{F}_m = \frac{f_m - f_m^{min}}{f_m^{max} - f_m^{max}} \tag{3.19}$$

Coefficients of variations of root-mean-square energy

Root-mean-square energy (RMS) is computed in frequency domain, using Parsevals theorem. It is defined as

$$RMS = \sqrt{\sum_k |\frac{X(k)}{K}|^2} \tag{3.20}$$

Then, coefficients of variations can be calculated as

$$CVR = \sqrt{E(RMS - E(RMS))^2} \tag{3.21}$$

where E(X) denotes the average of X.

Multi-stage average spectrum

Multi-stage average spectrum (MSAS) is used to extract both the time-varying and frequency-varying features within each frog syllable. The detailed steps for the MSAS method are described as follows:

Step 1: Divide all the frames of a frog syllable into 3 time stages equally. Then, reclassify the similar frames into the same stage in accordance with the forward direction of time. For example, if a frame j is classified into stage i, the frame after frame j can only be classified into stage i or the stage after i.

Step 2: Calculate the distance between the spectrum of each frame and the averaged spectrum of each stage.

$$d(i,j) = \sqrt{\sum_{k=0}^{L-1}[X_j(k) - S_i(k)]^2} \tag{3.22}$$

where $d(i,j)$ is the distance between the j-th frame and the i-th stage, $S_i(k) = \sum_(j = 0)^(N_i - 1)\frac{|X_j(k)|}{N_i}$, $0 \leq k \leq N-1$, $|X_j(k)|$ is the amplitude spectrum of the j-th frame, N_i is the total number of frames in i-th stage.

Step 3: Calculate the shortest accumulation distance from the first to the last frame and stage. The initial point (1,1) denotes the distance between the first frame and the first stage. Let $Acc(j,i)$ denotes the shortest accumulation distance from the initial point to point (j,i) which is defined as.

$$Acc(i,j) = minAcc(j-1,i) + d(j,i), Acc(j-1,i-1) + d(j,i) \tag{3.23}$$

where $Acc(1,i) = d(1,j)$ for all i, the $minX$ operation denotes the minimum value of X. $Acc(j-1,i) + d(j,i)$ and $Acc(j-1,i-1) + d(j,i)$ are the two candidate paths, a and b, respectively, which are considered in the calculation of the shortest accumulation distance at the point (j,i). Since one frame can only be classified into one stage, the path from point $(j,i-1)$ to (j,i) is forbidden in this study.

Step 4: Search for the shortest path from the final point $(N_f, 3)$ back to the initial point (1,1) along with those paths that have a shorter distance between the two candidate paths after all of the shortest accumulation distances for each frame j and each stage i have been calculated. The points on the shortest path are recorded as a sequence of coordinate $P = P(1), P(2), , P(N_f)$, where N_f is the total number of frames in a syllable.

Step 5: Reclassify each frame into a new stage according to the coordinate sequence of the shortest path.

Step 6: Repeat steps 2 to 5 until the change in the updated shortest accumulation distance of

the final point is lower than the predefined threshold.

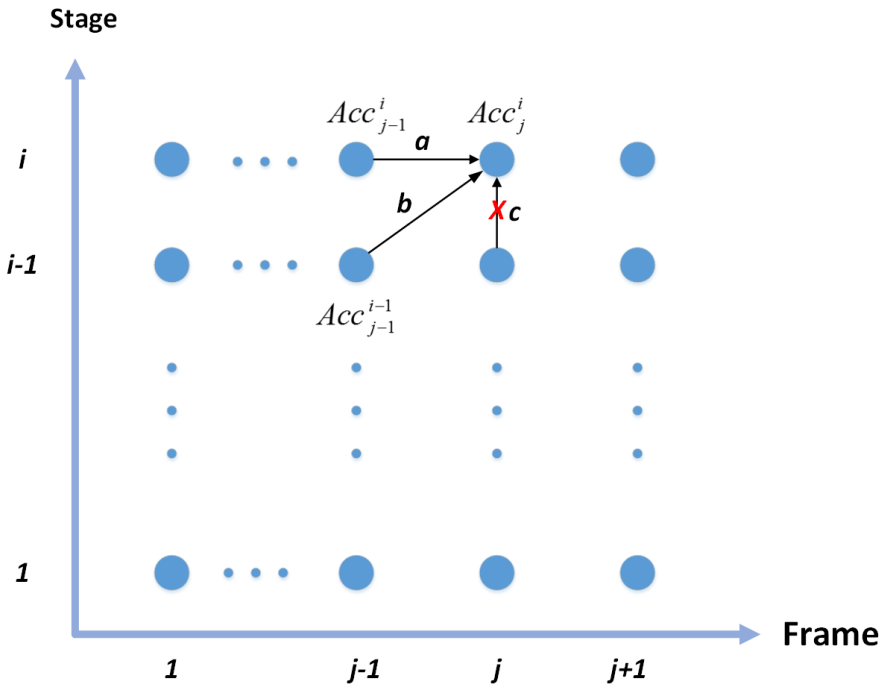


Figure 3.2: A frame and stage plane for calculating the shortest accumulation distance.

3.2 Classification

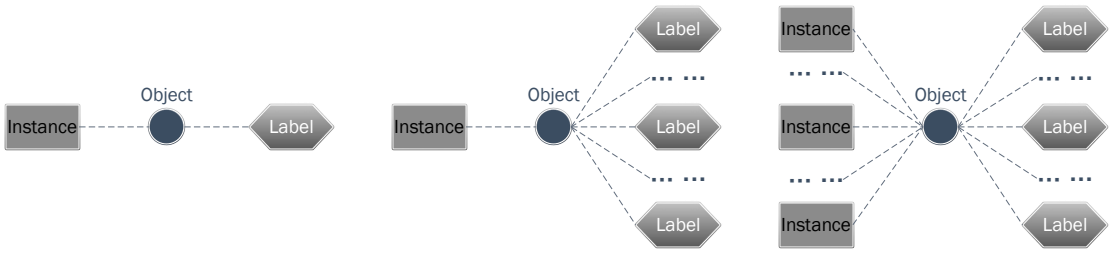
After feature extraction, the next step is to perform classification. This section aims to discuss various classification approaches used throughout this dissertation for classification of frog vocalisations.

3.2.1 Introduction

Recent advances in acoustic sensors have made it possible to collect acoustics data over large spatial and temporal scales. Also, large volumes of acoustic data are collected. Therefore, it is unavoidable to develop automatic means for data analysis. Over the past decades, machine learning has become a hot topic in the field of data analysis. Among various machine learning techniques, classification is one methodology to assign a class label to a set of unknown objects on the basis of a training data with known class membership, which is potentially useful for tagging collected acoustic data. For a classification process, it often consists of two stages: training and testing. In the training stage, a model is constructed using the testing data with

known class label. In the testing stage, the learned model is used to predict the class label for the testing data.

In previous studies, it is often assumes that there is only one frog species in one individual recording. Then, frog call classification is framed as a single-instance single-label (SISL) classification problem. Each segmented frog syllable (instance) is represented by a frog species (label). However, classify frog calls in environmental recordings can not fit in this framework well. In particular, each individual environmental recording consists of multiple simultaneously vocalising frog species and therefore multiple labels need to be assigned to one recording. To address this problem, two different classification frameworks are first introduced to study frog calls: multiple-instance multiple-label (MIML) classification and multiple-label (ML) classification. Compared to the traditional SISL classification framework, both MIML and ML classification frameworks are more convenient and natural for representing the frog species in environmental recordings. The frameworks for the SISL learning, ML learning and MIML learning are shown in Figure. 3.3.



(a) SISL learning (b) ML learning (c) MIML learning

Figure 3.3: Three different learning frameworks for frog call classification

The SISL classification can be generally divided into two types: binary classification and multi-class classification. For the binary classification, the data is separated into two classes. If the data is separated into one of several classes, then it is named as multi-class classification. Most classification methods have been specifically developed for binary classification. To extent the binary classification into multi-class classification, the combined use of multiple binary classifiers can be a solution. For a specific classification task, it is common that several classifiers are tested in order to find a suitable classifier with high classification accuracy.

For the ML classification, there are mainly two methods for tackling, which are problem transformation methods and algorithm adaptation methods. Problem transformation methods address the classification problem using single-class classifiers by transforming the ML problem

into a set of binary classification problems. In contrast, algorithm adaptation methods try to handle the ML classification problem in its full form. Therefore, the algorithms are directly applied to perform ML classification instead of simplifying the problem.

MIML classification is often addressed by using the multiple-instance (MI) learning or ML learning as the bridge.

3.2.2 SISL learning

Linear discriminant analysis (LDA)

LDA is a classification method originally developed in 1936 by R. A. Fisher [Fisher, 1936]. The goal of LDA is to improve the classification accuracy at a relatively low-dimensional feature space. To achieve this goal, the within-class distance needs to be minimised while the between-class distance needs to be maximised. In LDA, a feature space from n-dimension to d-dimension can be determined by using an optimal transformation matrix, where the transformation matrix is a linear mapping that maximised the Fisher criterion.

$$J_F(\omega) = \frac{\omega^T S_B \omega}{\omega^T S_W \omega} \tag{3.24}$$

where S_W and S_B are the between-class scatter matrix and within-class scatter matrix, respectively.

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (X_i^j - \mu_j)(X_i^j - \mu_j)^T \tag{3.25}$$

$$S_B = \sum_{j=1}^{C} (\mu_j - \mu)(\mu_j - \mu)^T \tag{3.26}$$

where X_i^j is the i-th vector of class j, μ_j is the mean vector of class j, C is the number of classes, and N_j is the number of feature vector in class j, μ is the mean vector of classes. Since LDA aims to find a transformation matrix that maximises the ratio of between-class scatter to within-class scatter in a lower-dimensional space. The optimal solution can be described as

$$\omega_{opt} = argmax_{\omega} \frac{\omega^T S_B \omega}{\omega^T S_W \omega} \tag{3.27}$$

where transformation matrix ω_{opt} can be determined by finding the eigenvector of $S_W^{-1} S_B$. In the recognition stage, features are first transformed into a lower-dimension by the transformation matrix ω_{opt}, derived by LDA. Then the distance between the feature vector of the test data and the feature vector representing the specific class is calculated. Lastly, the one with minimum distance is regarded as the recognised class.

K-nearest neighbour (K-NN)

K-NN is a non-parametric classification method, which has been widely used for animal call classification [Han et al., 2011, Huang et al., 2009, Xie et al., 2015b, 2016]. An object is classified to the class of majority of its k nearest neighbours [Huang et al., 2009]. Specifically, feature vectors are stored with class labels. For the test phase, the distances between an input feature vector and all stored vectors are calculated. Then, k closest vectors are used for selecting the most frequent vector as the class label. The distance function is defined using L^p norm, $p \in [1, \infty]$. Here the special cases are the Manhattan distance ($p = 1$), the Euclidean distance ($p = 2$), and the maximum distance ($p = \infty$). The distance metric is defined as

$$d_j^d = ||x - x^j||_p = (\sum_{i=1}^{d} |x_i - x_i^j|)^{1/p} \tag{3.28}$$

where i is index of the feature vector, j is the index of stored feature vector, d denotes the dimension of the input feature vector. Next, k nearest neighbours of the feature vector i is selected based on the distance for selecting the most frequency vector as the class label. For instance, if the following equation satisfied

$$\frac{1}{k_1} \sum_{j \in s_1} d(i, j(s_1) \leq \frac{1}{k_2} \sum_{j \in s_2} d(i, j(s_2) \tag{3.29}$$

where $k = k_1 + k_2$, k_1 is the number of class s_1, k_2 is the number of class s_2. Here the input feature vector i will be classified as class s_2.

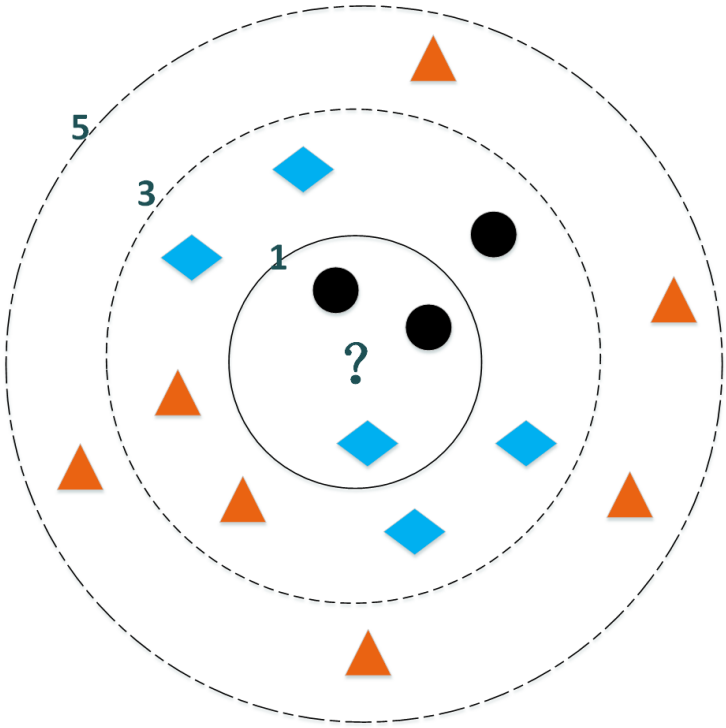


Figure 3.4: An example of K-NN algorithm process. If k is 1, then black circle will be assigned to the question mark; if k is 3 then the blue diamond will be assigned to the question mark; if k is 5, then green triangle will be assigned to the question mark

Decision tree (DT)

Decision tree is a tree structure that classifies instances by sorting them based on the values of feature vectors. It aims to create a model that predicts the class label of a target feature based on several input feature vectors. Compared to other classifiers, decision tree can provide a visual representation of the classification process. Also the importance of different feature vectors is given. An example of a decision tree is shown in Figure. 3.5. For a decision tree, each internal node is labelled with an input feature, and each leaf node is labelled with a class. To construct a decision tree, an attribute value test is recursively used to split the source feature set into subsets until the subset at a node has the same value of the target variable or splitting no longer adds value to the predictions. The most common strategy for learning decision tree from data is top-down induction of decision trees [Quinlan, 1986], which is an example of a greedy algorithm.

Given a set of training instances, a feature set and cut-off value τ are used to split the data into two sets: those instances for which $X_1 = x|x(j) \leq \tau$ and those for which $X_2 = x|x(j) \geq \tau$. The choice of feature j and value τ is made by minimising some measure of

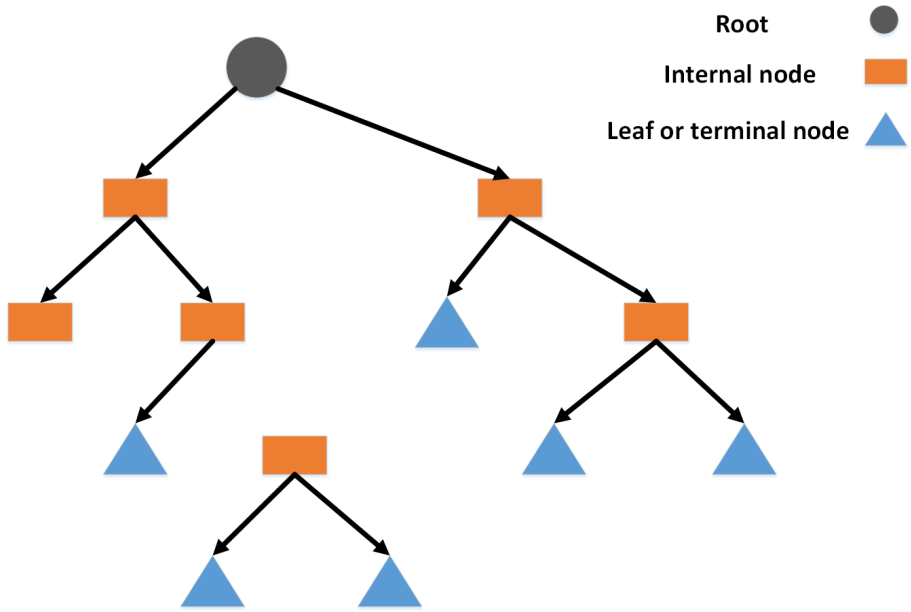


Figure 3.5: An example of decision tree structure with three components

impurity of sets X_1 and X_2 with respect to the classes of the points in each set. A commonly used measure is the Gini index, which is defined as

$$G(X) = \sum_{c \in C} p_c(1 - p_c) \tag{3.30}$$

where p_c is the proportions of instances in F of class c. The next is to find optimised feature j and cutoff τ to maximise $G(X) - (G(X_1) + G(X_2))$.

$$(j_{opt}, c_{opt}) = argmax_{j,\tau}(G(X) - (G(X_1) + G(X_2))) \tag{3.31}$$

where X is the current set of training instances. Once the set of instances are split into sets X_1 and X_2, decision tree are recursively built on X_1 and X_2 until some stopping criteria is met, usually a threshold of impurity or a threshold on the minimum number of training instances in a sub-tree.

TreeBagger

To improve performance of the decision tree algorithm and avoid over-fitting, a general technique of bootstrap aggregating or bagging is applied to tree learners [Breiman, 1996]. There are two steps for this process: training and prediction. In the training stage, a random sample

with replacement of training set is repeatedly selected. Then those samples are used to train different decision trees. After training, predictions for unseen samples are made by taking the majority vote in the case of decision trees. The advantage of this bootstrapping procedure is that it decreases the variance of the model without increasing the bias.

From bagging to random forest

Based on the above bagging technique, random forest selects a random subset of the features at each candidate split in the learning process, which is sometimes called Feature bagging [Ho, 1995]. The reason for doing this is the correlation of the trees in an ordinary bootstrap sample.

Support vector machines (SVM)

The support vector machine algorithm [Cortes and Vapnik, 1995] often constructs a hyper-plane or set of hyper-planes in a high or infinite-dimensional space to perform classification tasks. To achieve a good classification performance, the hyper-plane is computed to have the largest distance to the nearest training-data point of any classes (maximal marginal). For a p-dimension data, a $(p-1)$-dimension hyper-plane is sought to separate. To explicitly explain the ideal of maximal marginal, suppose that the task is to classify data points each belong to one of two classes as shown in Figure.3.6. To separate the classes, there are an infinite number of lines such as $H(i), i = 2, , N-1$, but how to find the best choice ($H(N)$) based on some pre-defined requirements is the question. The reason for choosing the best line (maximal marginal) is that the one represents the best separation. An illustration of choosing the maximum margin for data in two dimensions is given in Figure. 3.7. The boundaries on each side are defined by dashed lines, where the region between them is called the margin. Support vectors are defined as those data points on the boundaries. Those data points are closest to the decision boundary (separating hyper-plane) and play the most important role in determining the margin by which are two classes are distinguished.

Let $x_i, i = 1, \ldots, M$ represent the training input for a binary classification task, $C_i \in -1, +1$. For the linearly separated data, the hyper-plane can then be defined as

$$D(x) = w^T x + b \tag{3.32}$$

where w is the weighting vector perpendicular to the hyper-plane, b is the scalar bias.

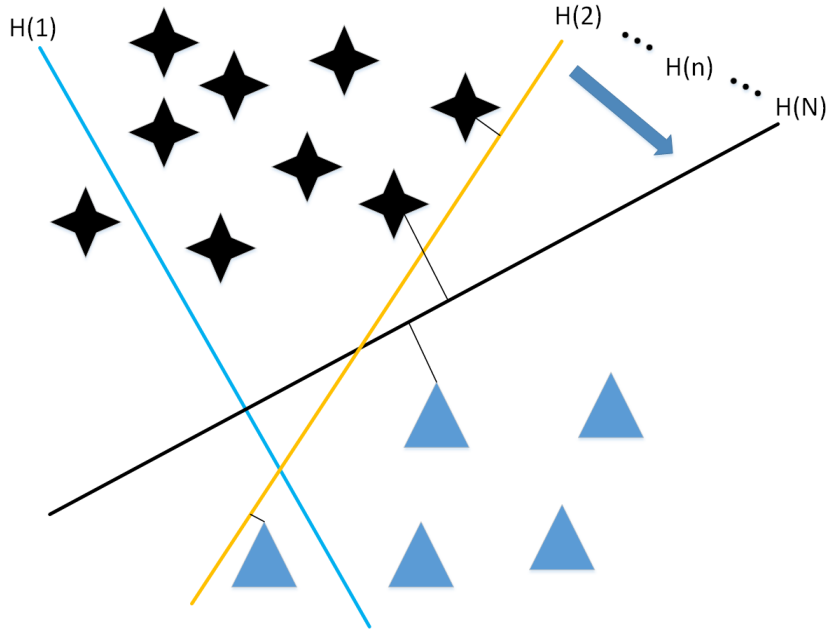


Figure 3.6: Linear separation of data for a two-class problem. Here $H(1)$ does not separate the classes; $(H(i)|i = 2 : N-1)$ does, but only with a small margin; $H(3)$ separates them with the maximum margin.

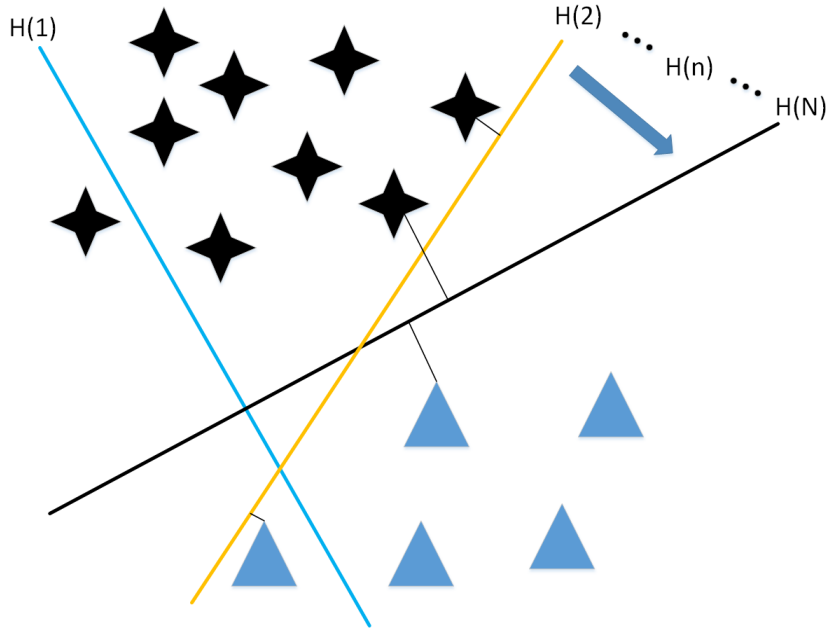


Figure 3.7: Linear separation of data using the maximum margin criteria

The distance between the dashed boundary planes is inversely proportional to $||w||$ and the optimisation problem is equivalent to minimising the function

$$Q(w, b) = 1/2||w||^2 \tag{3.33}$$

Subject to the constrains that the training data satisfy

$$y_i(w^T x_i + b) \geq 1, i = 1, , M \tag{3.34}$$

It is possible to solve the problem with a quadratic programming algorithm when $||w||^2$chooses the Euclidean norm. A function of the support vectors that satisfy Eq. 3.34 is used as the optimal hyper-plane. And it is adapt to the new training data. Therefore using standard optimisation procedures may be able to solve the minimisation problem.

When addressing soft-margin problems where training samples are linearly inseparable, a non-negative slack variable ξ is added to constraints and the object function. If $0 < \xi < 1$, the maximum margin does not exist in the training data. Also the aforementioned linear discriminant function may not work well for some data which in real application are corresponding features in feature space. So a non-linear kernel is introduced to map the data into a higher dimensional space by following Mercers theorem [Burges, 1998]. Polynomial kernel, Gaussian kernel, sigmoid kernel are some examples frequently used by researches. The SVM algorithm is usually known as the binary classifier. A conventional remedy to make SVM possible for multi-class classification is the so called one-against-all method, which separates one class from other classes at each time. Therefore, to classify M classes, the SVM algorithm needs to be run M times.

Artificial neural network (ANN)

Artificial neural network (ANN) is a non-linear, adaptive, machine learning tool built on connection principles [Lek and Guégan, 1999, Samarasinghe, 2006]. Because of its massively parallel and distributed structure, ANN has great capabilities for learning, generalization, non-linear approximation, thus for classification. For a linear model, it is based on linear combinations of fixed non-linear basis functions $\phi_j(X)$ such as

$$y(X, W) = f(\sum_{j=1}^{W} w_j \phi_j(X)) \tag{3.35}$$

where $f(\cdot)$ is non-linear activation function in the case of classification.

For neural networks, each basis function is itself a non-linear combination of a linear combination of the inputs, where the coefficients in the linear combination are adaptive parameter.

The basis neural network model can be described as a series of functional transformations. First we construct M linear combinations of the input variables $x_1, \ldots, x_D$ in the form

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(0)} \tag{3.36}$$

where $j = 1, \ldots, M$, and superscript (1) indicated the corresponding parameters are in the first 'layer' of the network. The parameters $w_{ji}^{(1)}$ and $w_{j0}^{(0)}$ are used as *weights* and *biases*, respectively.

The quantities a_j are known as *activations*. Then each of them is transformed using a differentiable, nonlinear activation function $h(\cdot)$ to give

$$z_j = h(a_j) \tag{3.37}$$

These quantities correspond to the outputs of the basis functions in Equation (3.35), which are called $hidden\ units$ in the context of neural network. The non-linear functions $h(\cdot)$ are generally chosen to be sigmoid functions such as the logistic sigmoid or the 'tanh' function. Then, the values are again linearly combined to give $output\ unit\ activations$

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{3.38}$$

where $k = 1, \ldots, K$, and K is the total number of outputs. This transformation corresponds to the second layer of the network, and again the $w_{k0}^{(2)}$ are bias parameters. Finally. the output unit activations are transformed using an appropriate activation function to give a set of network outputs y_k. The choice of activation function is determined by the nature of the data and the assumed distribution of target variables. For multiple binary classification problems, each output unit activation is transformed using a logistic sigmoid function so that

$$y_k = \sigma(a_k) \tag{3.39}$$

where

$$\sigma(a) = \frac{1}{1 + exp(-a)} \tag{3.40}$$

Combining various stages, the overall network function for sigmoid output uni activation functions can be represented as

$$y_k(X, W) = \sigma(\sum_{j=1}^{M} w_{kj}^{(2)} h(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}) + w_{k0}^{(2)}) \tag{3.41}$$

where the set of all weight and bias parameters have been grouped together into a vector W. Thus the neural network model is simply a non-linear function from a set of input variables x_i to a set of output variables y_k controlled by a vector W of adjustable parameters.

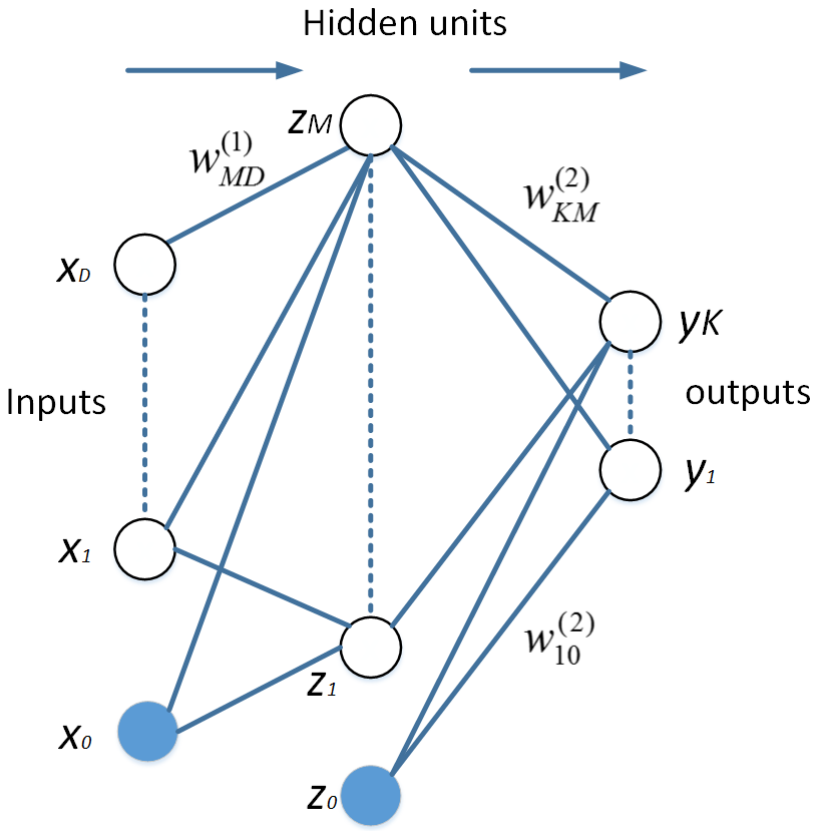


Figure 3.8: Network diagram for the two-layer neural network. The input, hidden, and output variables are represented by nodes, and the weight parameters are presented by links between the nodes. The bias parameters are denoted by links coming from additional input and hidden variables x_0 and z_0. Arrows denote the direction of information flow through the network during forward propagation

3.2.3 MIML learning

MIML-KNN

MIML-SVM

MIML-RBF

3.2.4 ML learning

Binary relevance for ML classification

Classifier chains for ML classification

Random k-labelsets for ML classification

3.3 Conclusion

Various feature extraction methods and classifiers (single-instance single-label) that have been used in previous studies are described. The purpose of feature representation is to represent frog vocalisations into a feature vector, which is effective enough to separate frog calls from the background noise. In high SNR recordings, most features can achieve good classification performance. However, the classification performance decreases rapidly for the environment recordings. For the classifiers, they are often designed for the classification of frog vocalisations. Since previous studies often assume that there is only one frog species in each individual recording, SISL learning is used for classification. Unfortunately, most environment recordings consists of more that one frog species. Therefore, it is worth to employ new classification frameworks to classify environment recordings with multiple simultaneously vocalising frog species.

Chapter 4

Frog call classification based on enhanced features and machine learning algorithms

---

4.1 Introduction

This chapter presents an enhanced feature representation for the frog call classification using various machine learning algorithms. In the literature, various feature representations have been developed for frog call classification. However, most features used in the prior work are based on either temporal features, perceptual features, or cepstral features. Is it 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?

This chapter aims to compare various feature representations with different machine learning techniques. Based on the classification performance, suggested features can then be transplanted to study low-SNR recordings. This chapter directly addresses sub-research question 1(a): Various acoustic features have been developed to classify frog calls in high SNR recordings; which of those features can be transplanted from high SNR recordings to low SNR recordings?

The performance of the proposed method is evaluated based on twenty-four frog species, which are geographically well distributed through Queensland, Australia. Five feature representations are compared via five machine learning algorithms. Classification results demonstrate a combination of temporal, spectral and cepstral features can achieve the best performance.

Compared with temporal and spectral features, cepstral features can achieve a robust classification accuracy, but sensitive to the background noise. Therefore, in the next chapter, we aim to develop a robust cepstral features that are not sensitive to the background noise.

4.2 Journal paper - Acoustic classification of Australian frogs based on enhanced features and machine learning algorithms

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

Email address: j3.xie@student.qut.edu.au

{*m.towsey, jinglan.zhang, p.roe*}*@qut.edu.au*

Abstract

Frogs are often considered as excellent bio-indicators, but a steady decrease in frog population has been noticed worldwide. To monitor the change of frog population and optimise the protection policy, frog call classification becomes an important topic. However, automatic classification of frog calls has not been adequately addressed in the literature. In this paper, classifying frog calls based on the enhanced temporal, perceptual and cepstral (*TemPerCep*) features is proposed. Time-frequency information of frog calls can be effectively represented via the enhanced *TemPerCep* feature, which gives a good classification performance. To be specific, continuous frog recordings are first segmented into individual syllables. Then, 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 various machine learning algorithms to perform frog call classification. Twenty-four frog species, which are geographically well distributed throughout Queensland, Australia, are used in this experiment. Experiment results show outstanding performance of the proposed feature representation, compared with the baselines.

Keywords: Frog call classification, Feature fusion, Support vector machine, Random forest, Artificial neural network

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 Quinlans 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

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

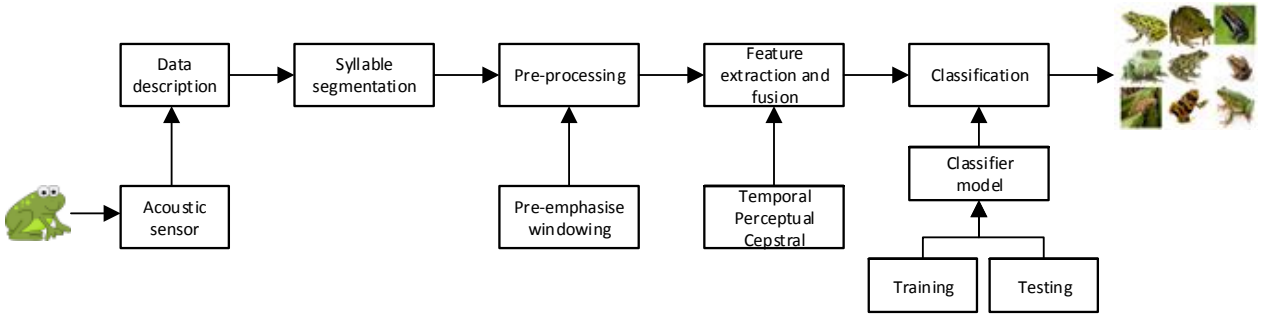


Figure 1: Flowchart of frog call classification system

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 preclassification 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 previsous 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 various 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 refereed 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 the continuous multiple calls of one frog species. For frogs, one syllable is an elementary acoustic unit for classification, which is a continuous frog vocalization emitted from an individual [7]. In this study, one method built on H*ä*rm*ä*'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

Table 1: Summary of scientific name, common name, and corresponding code. Frog species name with asterisk means 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

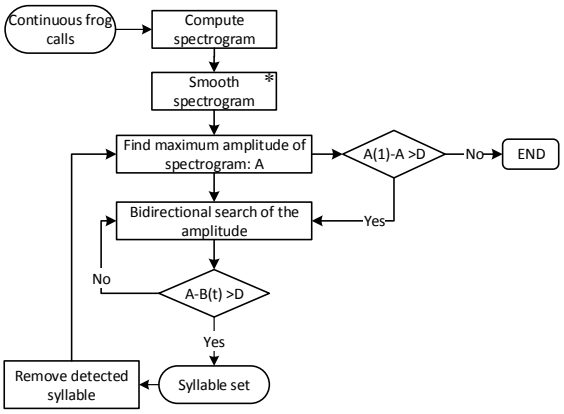


Figure 2: Segmentation method based on Härmä's algorithm. Here, D is the amplitude threshold for stopping criteria which is set at 20 dB experimentally, and the segmentation result is sensitive with this value. A is the maximum amplitude value of the spectrogram and we save the first maximum amplitude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

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 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 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 end. To enhance those high-frequency components and reduce the low-frequency components, 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 n-th sample of the syllable.

2.3.2. Windowing

After pre-emphasise, 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

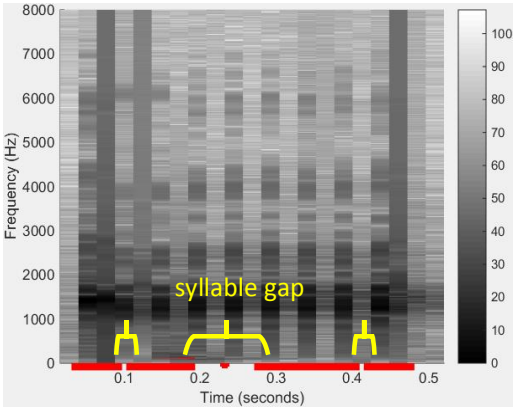
$$w(n) = 0.54 - 0.46cos(\frac{2n\pi}{L-1}), 0 \le n \le L-1 \tag{2}$$

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 window process is expressed as

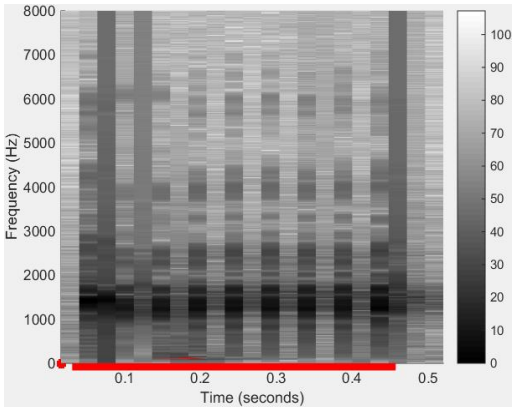
$$x(n) = w(n)y(n) \tag{3}$$

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]. Also, MFCCs achieved a better performance than LPC [11]. Different from hybrid features used in [7, 9, 3, 14], our enhanced enhanced



(a) Segmentation result without smoothing



(b) Segmentation result with smoothing

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

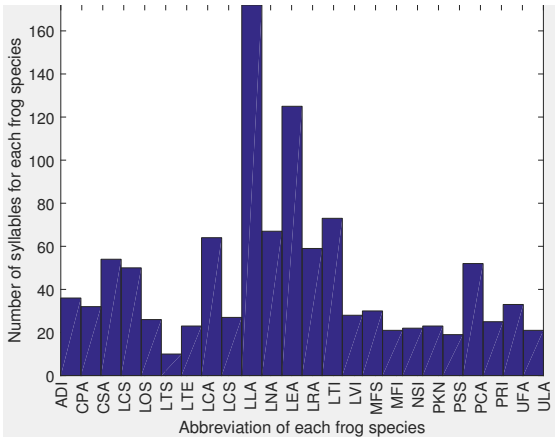


Figure 4: Distribution of syllable number for all frog species. The x-axis is the abbreviation of each frog species, and the corresponding scientific name can be found in Table 1.

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. MFCCs feature is used as a cepstral feature. The description of each feature is list 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 legnth 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

$$Re = \frac{1}{1-\alpha} log_2(\sum_{i}^{n} p_i^{\alpha}) \tag{6}$$

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

(3) 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

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

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

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

(5) 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 uncertain. 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].

(6) 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 k-th frame, N is the half size of DFT.

(7) 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 of the between the geometric mean and the arithmetic mean of the power spectrum and defined as

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

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

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

where θ is set at 0.85.

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

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

(10) 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 half-wave rectifier function.

(11) 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.

(12) 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).

(13) 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.

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

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

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

Step 2: Discrete cosine transform: MFCCs for the i^{th} frame are computed by performing DCT on the logarithm of E_j.

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

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.

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

where f_m is the m^{th} 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 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 we report the classification results for five machine learning algorithms: 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 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

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

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

where F_i^j is the i-th 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 $S_W^{-1} S_B$.

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

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

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

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 generalization 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 $(F_l^n, L_l^n), l = 1, 2, ..., C_l$ 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 *l-th* 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.

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

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 n-th 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, generalization, 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 used 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.

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

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.

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

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

where $i = 1, ..., N$, $j = 1, ..., m_i$, 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 five-fold 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 than RF achieves the best classification performance, while the classification performance of LDA is the lowest. Meanwhile, the classification performances 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 are very high. It can also be noted that RF can be highly recommended for classification of frog calls due to the highest classification accuracy.

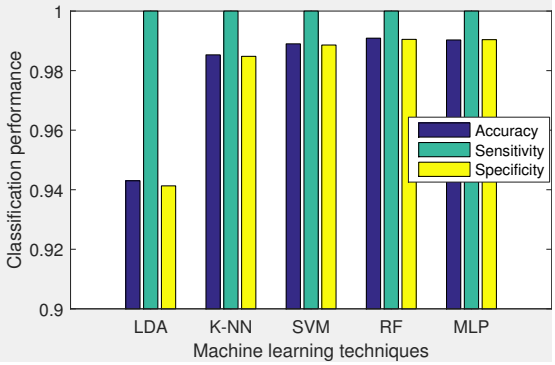


Figure 5: Classification results of different machine learning algorithms

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 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 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 different

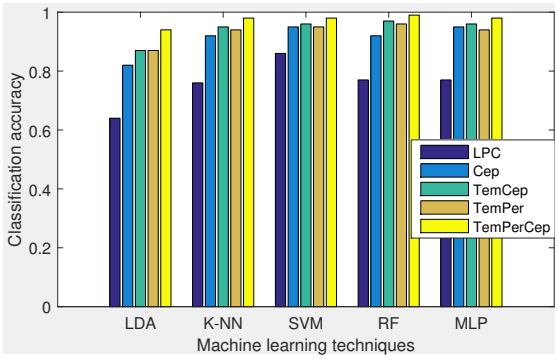


Figure 6: Classification results with different feature representations

classification performance (Fig. 7 and Fig. 8). The window sizes used for test are 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 that temporal features have a high weight in *TemPer* for the classification task.

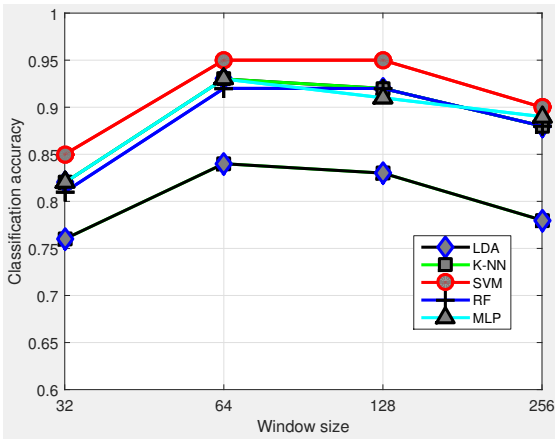


Figure 7: Classification results of MFCCs with different window sizes

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 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

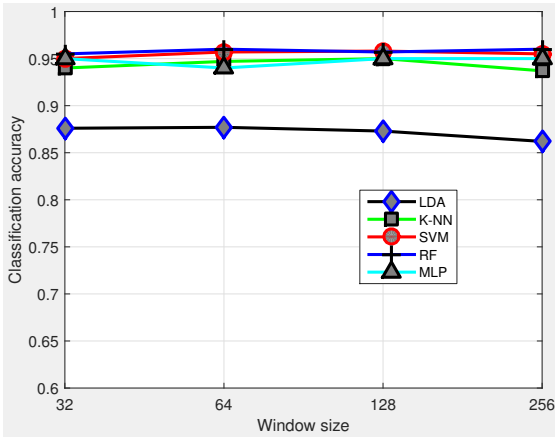


Figure 8: Classification results of *TemPer* with different window sizes

of noise contamination are shown in Fig. 9. It is found from Fig. 9 that MFCCs (Cep) are very sensitive to the 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 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.

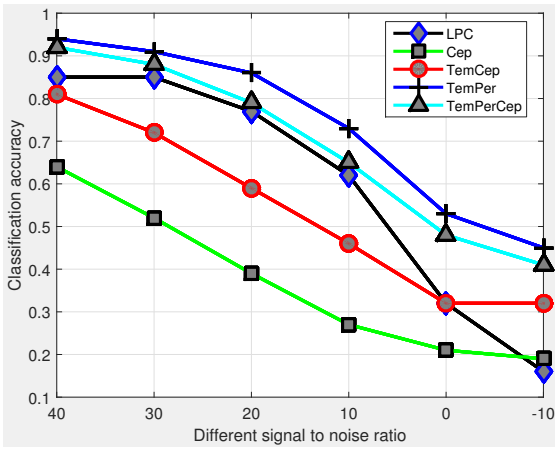


Figure 9: Sensitivity of different feature representations for different levels of noise contamination

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 our dataset with the same classifier (SVM). Compared with those previous methods, our proposed enhanced feature representation significantly outperforms other methods. Therefore, we can conclude that our feature representation can effectively characterise different frog calls. From the

Table 2: Comparison with previous used feature representations

Ref.	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%

Table 2, we can also observe that MFCCs is the most popular feature that have 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 perceptrual 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 outperforms 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.

In the future, since 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 studying frog calls. Also, a wider variety of frog audio data from different geographical and environmental conditions will be test in the future experiments.

6. Acknowledgement

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).

7. Reference

[1] J. Wimmer, M. Towsey, B. Planitz, I. Williamson, P. Roe, Analysing environmental acoustic data through collaboration and automation, Future Generation Computer Systems 29 (2013) 560–568.

[2] K. D. Wells, The ecology and behavior of amphibians, University of Chicago Press, 2010.

[3] B. Gingras, W. T. Fitch, A three-parameter model for classifying anurans into four genera based on advertisement calls, The Journal of the Acoustical Society of America 133 (2013) 547–559.

[4] J. Xie, M. Towsey, K. Yasumiba, J. Zhang, P. Roe, 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] G. Grigg, A. Taylor, H. Mc Callum, G. Watson, Monitoring frog communities: an application of machine learning, in: Proceedings of Eighth Innovative Applications of Artificial Intelligence Conference, Portland Oregon, pp. 1564–1569.

[6] C.-H. Lee, C.-H. Chou, C.-C. Han, R.-Z. Huang, Automatic recognition of animal vocalizations using averaged mfcc and linear discriminant analysis, Pattern Recognition Letters 27 (2006) 93–101.

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

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

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

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

[11] C. L. T. Yuan, D. A. Ramli, Frog sound identification system for frog species recognition, in: Context-Aware Systems and Applications, Springer, 2012, pp. 41–50.

[12] C.-J. Huang, Y.-J. Chen, H.-M. Chen, J.-J. Jian, S.-C. Tseng, Y.-J. Yang, P.-A. Hsu, Intelligent feature extraction and classification of anuran vocalizations, Applied Soft Computing 19 (2014) 1–7.

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

[14] J. Xie, M. Towsey, A. Truskinger, P. Eichinski, J. Zhang, P. Roe, 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] D. Stewart, Australian frog calls: subtropical east, Audio CD, 1999.

[16] A. Harma, Automatic identification of bird species based on sinusoidal modeling of syllables, in: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on, volume 5, IEEE, pp. V–545.

[17] H. Jaafar, D. A. Ramli, Automatic syllables segmentation for frog identification system, in: Signal Processing and its Applications (CSPA), 2013 IEEE 9th International Colloquium on, IEEE, pp. 224–228.

[18] F. Itakura, Line spectrum representation of linear predictor coefficients of speech signals, The Journal of the Acoustical Society of America 57 (1975) S35–S35.

[19] C.-H. Lee, C.-H. Chou, C.-C. Han, R.-Z. Huang, Automatic recognition of animal vocalizations using averaged mfcc and linear discriminant analysis, Pattern Recognition Letters 27 (2006) 93 – 101.

[20] W. Han, C.-F. Chan, C.-S. Choy, K.-P. Pun, An efficient mfcc extraction method in speech recognition, in: Circuits and Systems, 2006. ISCAS

2006. Proceedings. 2006 IEEE International Symposium on, IEEE, pp. 4–pp.

[21] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.

[22] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST) 2 (2011) 27.

[23] L. Bao, Y. Cui, Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information, Bioinformatics 21 (2005) 2185–2190.

[24] L. M. M. Juan Mayor, 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, 2009.

Chapter 5

Adaptive frequency scaled wavelet packet decomposition for frog call classification

---

5.1 Introduction

This chapter presents a novel cepstral feature representation based on adaptive frequency scaled wavelet packet decomposition. Following the conclusion of chapter 4 that cepstral features can achieve a good classification accuracy, but sensitive to the background noise, this research aims to develop a ceptral feature representation with a good anti-noise ability.

Since this thesis will address the low SNR recordings rather than high SNR recordings, developing feature representations with a good anti-noise ability is important. Different from most previous studies that extracted features using the Fourier transform, wavelet packet decomposition is employed in this study for feature extraction. The classification performance is evaluated with two different datasets from Queensland, Australia (18 frog species from commercial recordings (high SNR) and field recordings of 8 frog species from James Cook University recordings (low SNR)). This chapter answers the research question 2(b): How to improve the performance of developed features for frog call classification in low SNR recordings?

Although low SNR recordings are used in this research, we still regard the classification task as a single-instance single-label learning problem. However, most low SNR recordings have more than one frog species. Therefore, in the next two chapters, we focus on the classification of multiple frog species in one individual recording.

5.2 Journal paper - Adaptive frequency scaled wavelet packet decomposition for frog call classification

Ecological Informatics 32 (2016) 134–144

Contents lists available at ScienceDirect

Ecological Informatics

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

Adaptive frequency scaled wavelet packet decomposition for frog call classification

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

Electrical Engineering and Computer Science School, Queensland University of Technology, Brisbane, Australia

ARTICLE INFO

Article history:
Received 8 September 2015
Received in revised form 26 January 2016
Accepted 27 January 2016
Available online 4 February 2016

Keywords:
Frog call classification
Spectral peak track
Adaptive frequency scaled wavelet packet decomposition
k-means clustering
k-nearest neighbour
Support vector machine

ABSTRACT

Environmental changes have put great pressure on biological systems leading to the rapid decline of biodiversity. To monitor this change and protect biodiversity, animal vocalizations have been widely explored by the aid of deploying acoustic sensors in the field. Consequently, large volumes of acoustic data are collected. However, traditional manual methods that require ecologists to physically visit sites to collect biodiversity data are both costly and time consuming. Therefore it is essential to develop new semi-automated and automated methods to identify species in automated audio recordings. In this study, a novel feature extraction method based on wavelet packet decomposition is proposed for frog call classification. After syllable segmentation, the advertisement call of each frog syllable is represented by a spectral peak track, from which track duration, dominant frequency and oscillation rate are calculated. Then, a k-means clustering algorithm is applied to the dominant frequency, and the centroids of clustering results are used to generate the frequency scale for wavelet packet decomposition (WPD). Next, a new feature set named *adaptive frequency scaled wavelet packet decomposition sub-band cepstral coefficients* is extracted by performing WPD on the windowed frog calls. Furthermore, the statistics of all feature vectors over each windowed signal are calculated for producing the final feature set. Finally, two well-known classifiers, a k-nearest neighbour classifier and a support vector machine classifier, are used for classification. In our experiments, we use two different datasets from Queensland, Australia (18 frog species from commercial recordings and field recordings of 8 frog species from James Cook University recordings). The weighted classification accuracy with our proposed method is 99.5% and 97.4% for 18 frog species and 8 frog species respectively, which outperforms all other comparable methods.

1. Introduction

During the past decades, a rapid decline in frog biodiversity has been noted worldwide. There are many reasons for this decline, including habitat destruction (Clauzel et al., 2015), invasive species (Shine, 2014), and climate change (Garcia et al., 2014). Researchers investigate frogs to retain their biodiversity and develop effective protection strategies. Due to the development of acoustic sensor techniques, many sensors have been widely deployed for monitoring biodiversity, which produces large volumes of acoustic data (Wimmer et al., 2013). Compared with the traditional manual methods that require ecologists to physically visit sites for collecting biodiversity data, acoustic sensors can help collect audio data over larger spatio-temporal scales (Wimmer et al., 2010; Gage and Axel, 2014) Since several gigabytes of compressed data can be generated by an acoustic sensor per day, enabling automating species identification in acoustic data sets has become important (Zhang et al., 2013).

In recent years, acoustic data has been studied for the recognition and classification of animal calls by many researchers. Almost all the recognition and classification methods consist of four parts: pre-processing, syllable segmentation, feature extraction, and recognition or classification.

Frog call classification has been addressed in several papers. Huang et al. (2009) extracted spectral centroid, signal bandwidth, and threshold crossing rate from each segmented frog syllable. Then, two classifiers, k-nearest neighbour (k-NN) classifier and support vector machine (SVM), were used for classification. However, signal bandwidth and threshold crossing rate are very sensitive to the background noise, which results in low classification accuracy in noisy environments. Han et al. (2011) introduced spectral centroid, Shannon entropy and Rènyi entropy to classify frog calls with a k-NN classifier. Chen et al. (2012) first calculated syllable length for pre-classification of frog calls based on segmented frog syllables. Then, a multi-stage average spectrum was calculated for automatic recognition based on template matching. However, extracting features based on the Fourier transform has a tradeoff between time and frequency resolution, which restricts the discriminability of the features. Bedoya et al. (2014) proposed an automatic recognition system for frog calls based on the Mel-frequency cepstral coefficients (MFCCs) and a fuzzy classifier. However, MFCCs

* Corresponding author.
E-mail address: xiej8734@gmail.com (J. Xie).

http://dx.doi.org/10.1016/j.ecoinf.2016.01.007
1574-9541/

are designed for the human auditory system, and might be not suitable for the classification of frogs (Sahidullah and Saha, 2012). Meanwhile, MFCCs are not suitable for dealing with recordings with a low signal to noise ratio (SNR). In those previous studies (Huang et al., 2009; Han et al., 2011; Chen et al., 2012; Bedoya et al., 2014) most features used are either based on Fourier transform or transplanted from speech, speaker and music fields. To further improve the recognition and classification performance, it is necessary to develop more accurate species identification methods.

Wavelet analysis has been widely employed for acoustic data, because it can preserve both frequency and temporal information (Ren et al., 2008). Yen and Fu (2002) introduced wavelet packet transform (WPT) for individual frog identification. After applying WPT to the frog calls, energy of all the node coefficient were calculated as features. Then, Fisher's criterion (Yen and Lin, 2000) was used for dimension reduction. Finally, the feature vector after dimension reduction was fed into a neural network classifier for identification. Colonna et al. (2012) proposed to use discrete wavelet transform (DWT) for frog call classification. Based on the node coefficients of DWT, energy, power, zero-crossing rate and pitch of each node coefficients were calculated. However, applying WPT and DWT without any modifications cannot provide a good frequency domain resolution for classifying frog calls.

In this study, the WPD is applied to the frog calls with an adaptive frequency scale for feature extraction. Frog species that are genetically similar often share close advertisement calls (Gingras and Fitch, 2013). Therefore, the dominant frequency which is directly calculated from the trace of advertisement call is an important feature for differentiating frog species. We use dominant frequency to produce the frequency scale for WPD, which is different from using minimum and maximum frequency to generate the frequency scale for WPD in Ren et al. (2008). Specifically, continuous acoustic data are first segmented into syllables using Härmä's method (Harma, 2003). Then, spectral peak tracks are extracted from each syllable where possible. Three features are extracted from each track: track duration, dominant frequency and oscillation rate. Next, a k-means clustering algorithm is applied to the dominant frequency, and the centroids of clustering results are used to generate the frequency scale for WPD. After applying the adaptive frequency scaled WPD to the frog calls, a new feature set named *adaptive frequency scaled wavelet packet decomposition sub-band cepstral coefficients* (AWSCCs) is extracted. Finally, two classifiers, a k-NN classifier and a SVM classifier, are employed for the classification with the proposed feature set.

2. Methods

The architecture of the proposed classification method consists of five modules: syllable segmentation, syllable feature extraction, adaptive frequency scale generation, WPD feature extraction and classification (see Fig. 1). Each module is described in the following sections.

2.1. Sound recording and pre-processing

Two datasets obtained from a commercial recording (Stewart, 1999) and James Cook University (JCU) were selected for this study. Recordings, which were collected from the CD, are two-channel, sampled at 44.10 kHz and saved in MP3 format. All recordings were obtained with a directional microphone and have a high signal to noise ratio (SNR). Each recording includes one frog species, and has a duration ranging from twenty-one to fifty-four seconds. The calls of eighteen frog species recorded in Queensland, Australia were used to develop the detailed methodology. To reduce the subsequent computational burden, all recordings were re-sampled at 16 kHz per second, mixed to mono, and saved in WAV format.

The JCU recordings were obtained from Kiyomi dam (S 19°22′16.0′,E146°27′31.3″) BG creek dam (S19°27′1.23″,E146°24′5.65″) and Stony creek dam (S 19°24′07.0″,E146°2551.3) in Townsville, using Song Meter (SM2) (Xie, 2016). The recordings were stored on 16 GB SD cards in 64 kbps MP3 mono format and have a low

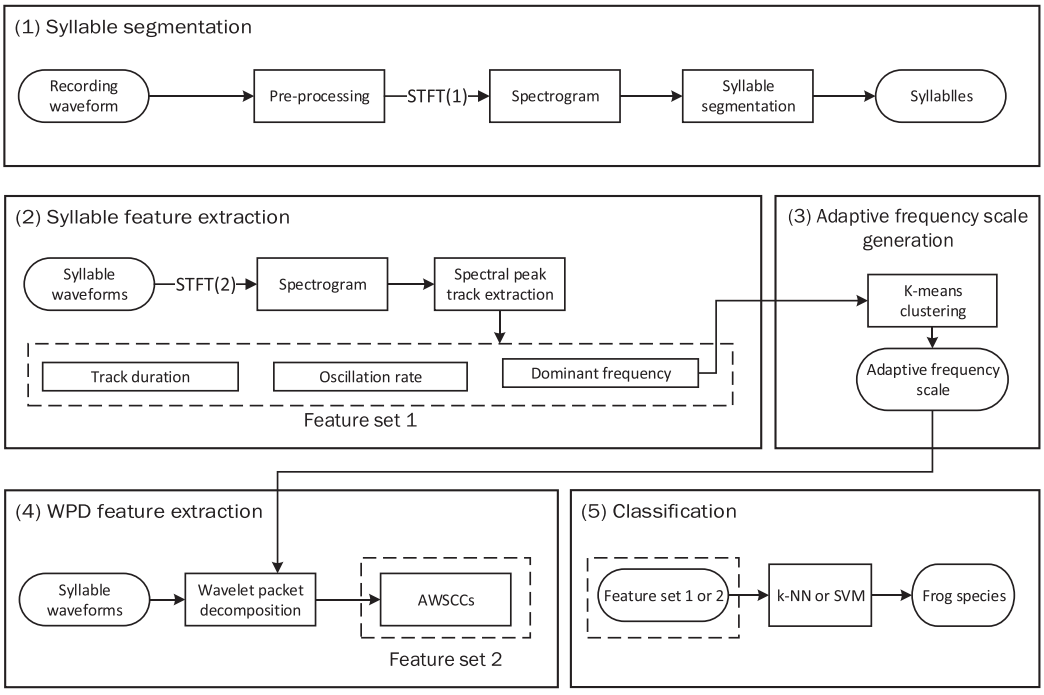


Fig. 1. Block diagram of the frog call classification system. The line of dashes indicates the extracted feature set. AWSCCs is the abbreviation of *adaptive wavelet packet decomposition sub-band cepstral coefficients*. STFT is short-time Fourier transform. For STFT(1), the window function, size and overlap are Kaiser window, 512 samples and 25%. For STFT(2), the window function, size and overlap are Hamming window, 128 samples and 90%. In this diagram, two feature sets are extracted, the description of other feature sets is shown in Fig. 6.

Table 1
Parameters of 18 frog species averaged of three randomly selected syllable samples in the commercial recording. These selected samples make the *reference data set*.

No.	Scientific name	Abbreviation	Syllable duration (millisecond)	Peak frequency (Hz)	Oscillation rate (cycle/s)
1	*Assa darlingtoni*	ADI	80	3200	160
2	*Crinia parinsignifera*	CPA	250	4300	350
3	*Litoria caerulea*	LCA	500	500	50
4	*Litoria chloris*	LCS	800	1700	220
5	*Litoria fallax*	LFX	430	4700	70
6	*Litoria gracilenta*	LGA	1400	2700	100
7	*Litoria latopalmata*	LLA	30	1400	2100
8	*Litoria nasuta*	LNA	100	2800	160
9	*Litoria revelata*	LRA	160	4100	70
10	*Litoria rubella*	LUA	500	2900	60
11	*Litoria verreauxii*verreauxii	LVV	270	3100	125
12	*Mixophyes fasciolatus*	MFS	200	1200	140
13	*Mixophyes fleayi*	MFI	50	1000	140
14	*Philoria kundagungan*	PKN	170	430	95
15	*Pseudophryne coriacea*	PCA	300	2400	80
16	*Pseudophryne raveni*	PRI	370	2500	45
17	*Rheobatrachus silus*	RSS	510	1500	60
18	*Uperoleia laevigata*	ULA	450	2400	150

SNR compared with the commercial recording. All the JCU recordings started around sunset, finished around sunrise every day and have 12 h duration.

2.2. Spectrogram analysis based on validation set

In this study, three syllables for each frog species are set aside and used as a *reference data set*. For the commercial recording, three parameters including syllable duration, dominant frequency, and oscillation rate, are manually calculated for those three syllables of each species and averaged, as listed in Table 1. The reference data set is excluded from the data used in the testing stage.

For the JCU recordings,[1] the corresponding parameters are described in Table 2. Compared with recordings from the commercial recording, peak frequency shows a smaller variation than syllable duration and oscillation rate.

2.3. Syllable segmentation

For frog calls, an elementary acoustic unit for classification is the syllable, which is a continuous vocalization emitted from an individual (Huang et al., 2009). Each commercial recording consists of the continuous multiple calls of one frog species. Therefore, it is necessary to segment each call into individual syllables. This syllable segmentation process is applied to the spectrogram, which is generated by applying short-time Fourier transform (STFT) to each recording. For STFT, the window function is the Hamming window with the size and overlap of 512 samples and 25%, respectively.

To further improve the segmentation result, those syllables whose averaged energy is less than 15% of the maximum energy are removed (Gingras and Fitch, 2013). The distribution of syllable numbers after segmentation for all frog species is shown in Fig. 2.

For the JCU recordings, bandpass filtering is applied to each recording before using Härmä's method. A bandpass filter is first used to filter specific frog species, because different frog species tend to call simultaneously. The filtering is

$$S'(t,f) = \begin{Bmatrix} S(t,f) & Fl_{ower} \le f \le F_{upper} \\ 0 & \end{Bmatrix}$$

[1] https://www.ecosounds.org/

Table 2
Parameters of 8 frog species obtained by averaging three randomly selected syllable samples from recordings of James Cook University. NA indicates there is no oscillation structure in the spectrogram for the background noise and frog chorus. Since syllable duration of *Rhinella marina* (commom name: Canetoad) is very different from each other, we manually set the duration of Canetoad using the maximum duration of other frog species, which is 500 ms.

No.	Scientific name	Abbreviation	Syllable duration (ms)	Peak frequency (Hz)	Oscillation rate (cycles/s)
1	*Rhinella marina*	CTD	500	680	12
2	*Cyclorana novaehollandiae*	CNE	350	600	NA
3	*Limnodynastes terraereginae*	LTE	80	630	NA
4	Litoria fallax	LFX	120	4100	50
5	Litoria nasuta	LNA	100	2700	NA
6	*Litoria rothii*	LRI	350	1150	15
7	Litoria rubella	LUA	500	2400	NA
8	Uperolela mimula	UMA	120	2400	40

Here, $S'(t,f)$ is the filtered spectrogram, the F_{lower} and F_{upper} are lower and upper cutoff frequency and calculated as

$$\begin{aligned} F_{upper} &= F_{peak} + \beta \\ F_{lower} &= F_{peark} - \beta \end{aligned} \quad (1)$$

where F_{peak} is the peak frequency (Table 2), β is a threshold for determining the frequency bandwidth and set at 300 Hz based on the *reference data set*.

After bandpass filtering, noise reduction is essential for improving the segmentation for the low signal to noise ratio in JCU recordings. Here, we use the method of Towsey et al. (2012) for noise reduction. Finally, we use Härmä's method to detect individual syllables (Fig. 3).

For *Canetoad*, the durations of different calls are very different, therefore, we manually selected 30 syllables whose combined duration is 500 ms.

For the JCU recordings, eight frog species were used for experiment. After syllable segmentation of continuous recordings, for each frog species, we randomly selected 30 syllables from segmentation results for subsequent analysis.

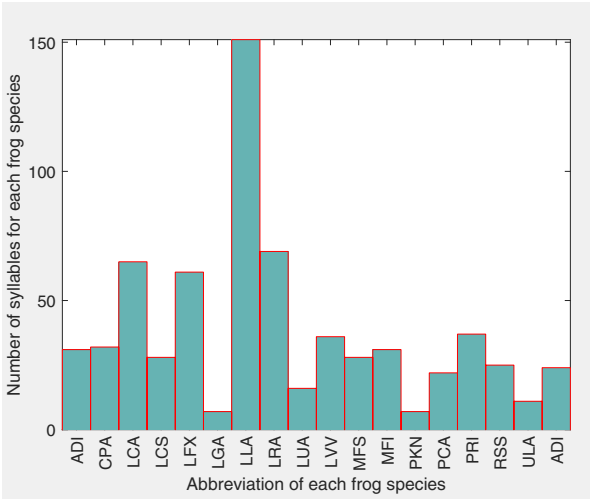


Fig. 2. Distribution of syllable number for all frog species. The x-axis is the abbreviation of each frog species, and the corresponding scientific name can be found in Table 1.

2.4. Spectral peak track extraction

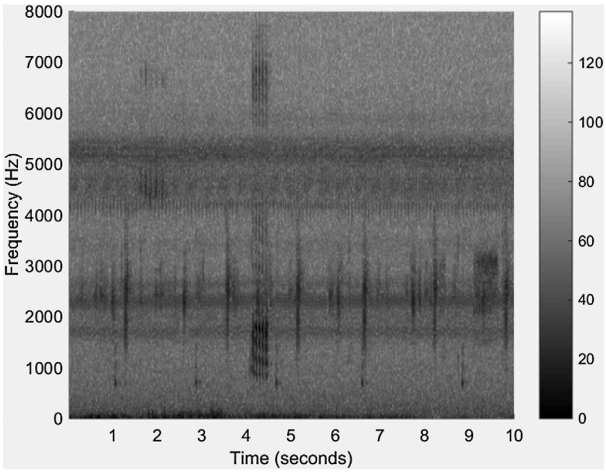
Spectral peak tracks (SPT) (also called frequency tracks) have been explored for studying birds (Jancovic and Kokuer, 2015; Heller and Pinezich, 2008) and whales (Roch et al., 2011). In this study, the spectral peak track is used to represent the trace of a frog advertisement call, because frogs which are genetically related share more similar advertisement calls than distantly related frogs (Gingras and Fitch, 2013). The reasons for using SPT are (1) to isolate the desired frog calls from the background noise; (2) to extract corresponding SPT features. Here, the SPT is extracted using a modified version of the method introduced in Xie et al. (2015) as follows.

For the SPT extraction algorithm, seven parameters need to be set (Table 3). The process for determining those parameters is explained in Section 3.

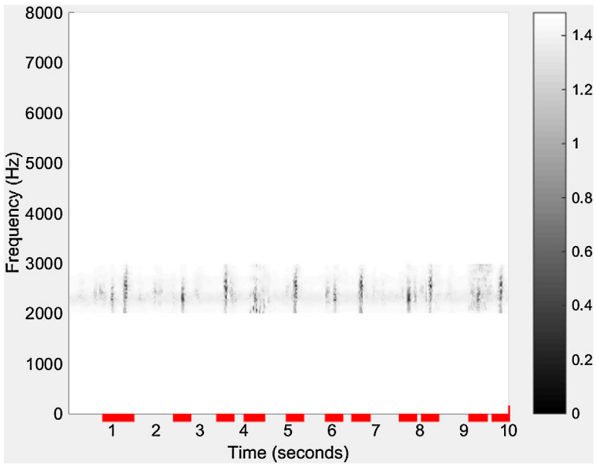
Before applying the SPT extraction algorithm, each syllable is transformed to a spectrogram with the following parameter settings (Hamming window, frame size is 128 samples, and window overlap is 90%). For the generated spectrogram, the maximum intensity (real peak) is selected from each frame with a minimum required intensity, I. Then, the time and frequency domain intervals between two successive peaks are calculated. If the time and frequency intervals are smaller than T_c and f_c respectively, one initial track (SPT_1) will be generated.

Table 3
Parameters used for spectral peak extraction.

Parameter	Description
I (dB)	Minimum intensity threshold for peak selection
T_c (s)	Maximum time domain interval for peak connection
T_s (s)	Minimum time interval for stopping growing tracks
f_c (Hz)	Maximum frequency domain interval for peak connection
d_{min} (s)	Minimum track duration
d_{max} (s)	Maximum track duration
β (0~1)	Minimum density value

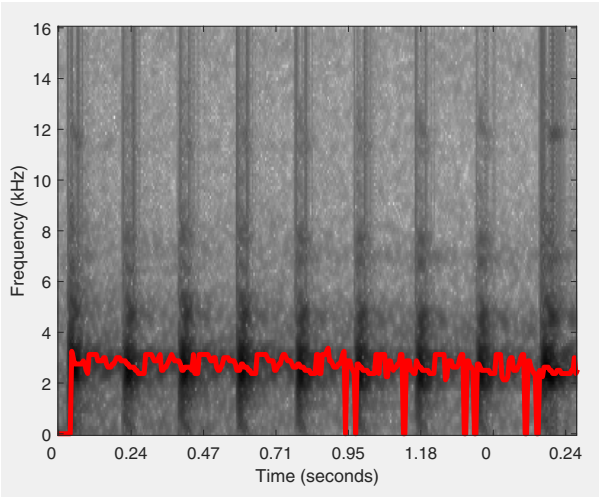


(a) Spectrogram.

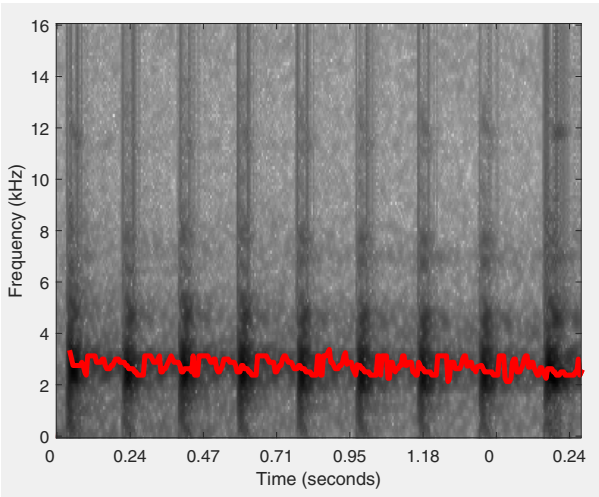


(b) Segmentation results with marked red lines.

Fig. 3. Segmentation results based on bandpass filtering for *Uperolela mimula*, noise reduction and Härmä's method. The red line in (b) indicates the start and stop location of each segmented syllable. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) selected peaks below the intensity threshold I and are set to zero.



(b) spectral peak track with predicted peaks using linear regression.

Fig. 4. Spectral peak track extraction results for *Neobatrachus sudelli*. By filling the gaps within the track, the dominant frequency can be more accurately calculated.

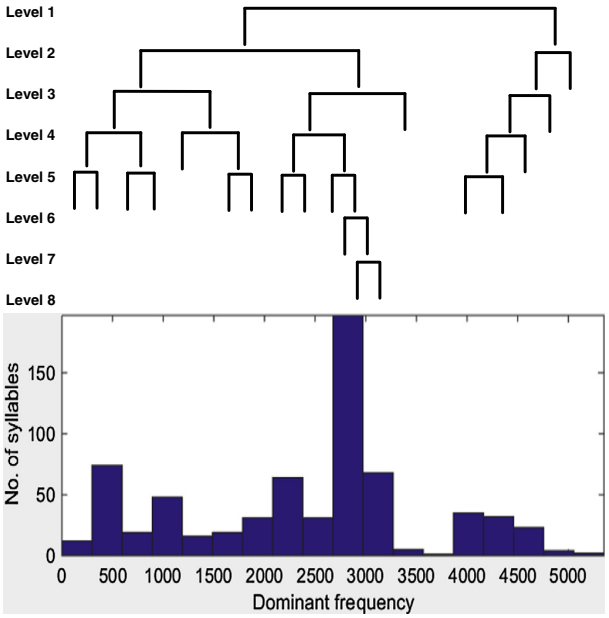


Fig. 5. Adaptive wavelet packet tree for classifying twenty frog species. The upper image is the wavelet packet tree; the lower image is the histogram of dominant frequency for twenty frog species.

After that, linear regression is applied to the generated track for calculating the position of the next predicted peak. Based on peaks $p_1(t_1,f_1)$ and $p_2(t_2,f_2)$ within the initial track (SPT_1), a and b in Eq. (2) can be solved.

$$f = at + b \tag{2}$$

Based on a and b, the predicted peak p_n of the following frame t_n can be calculated. Next, the time and frequency domain intervals between predicted peak (p_n) and the real peak of the successive frame are recalculated. If the time and frequency intervals are smaller than T_c and f_c respectively, the real peak will be added to the initial track. After each peak is added to the initial track, linear regression is repeated to recalculate the next predicted peak using at most the last 10 included peaks. This iterative process continues until T_s is no longer satisfied. When no more peaks will be added to one track, the next step is to compare the duration and density of the track with d_{min}, d_{max}, and β. If all conditions are satisfied, then the track will be saved to the track list. The SPT results for *Neobatrachus sudelli* are shown in Fig. 4. During the process of track extraction, time domain gaps are generated where the intensity threshold I is not reached. These gaps can be filled by predicting the correct frequency bin using linear regression, as illustrated in Fig. 4.

Table 4
Parameter setting for calculating spectral peak track.

Parameter	Commercial recordings	JCU recordings
I (dB)	3	3
T_c (s)	0.005	0.1
T_s (s)	0.05	0.2
f_c (Hz)	800	800
d_{min} (s)	0.01	0.05
d_{max} (s)	2	2
β (0 ~ 1)	0.8	0.6

2.5. Syllable SPT features

After SPT extraction, each SPT is expressed in the following format: (1) track start time t_s; (2) track stop time t_e; (3) frequency bin index for each of the peaks within the track f_t ($t_s \le t \le t_e$). Then, syllable features including track duration, dominant frequency, and oscillation rate are calculated based on the SPT.

(a) Track duration (second): Track duration (D_t) is directly obtained from the bounds of the track.

$$D_t = (t_e - t_s) * r_x \tag{3}$$

where r_x is the time domain resolution in unit second per frame.

(b) Dominant frequency (Hz): Dominant frequency ($\overline{f}$) is calculated by averaging the frequency of all peaks within one track

$$\overline{f} = \sum_{t=t_s}^{t_e} f_t/(t_e - t_s + 1) * r_y \tag{4}$$

where r_y is the frequency domain resolution with unit frequency per bin, f_t is the frequency bin index of peak t.

(c) Oscillation rate (Hz): Oscillation rate (O_r) represents the number of pulses per second. The algorithm for extracting oscillation rate

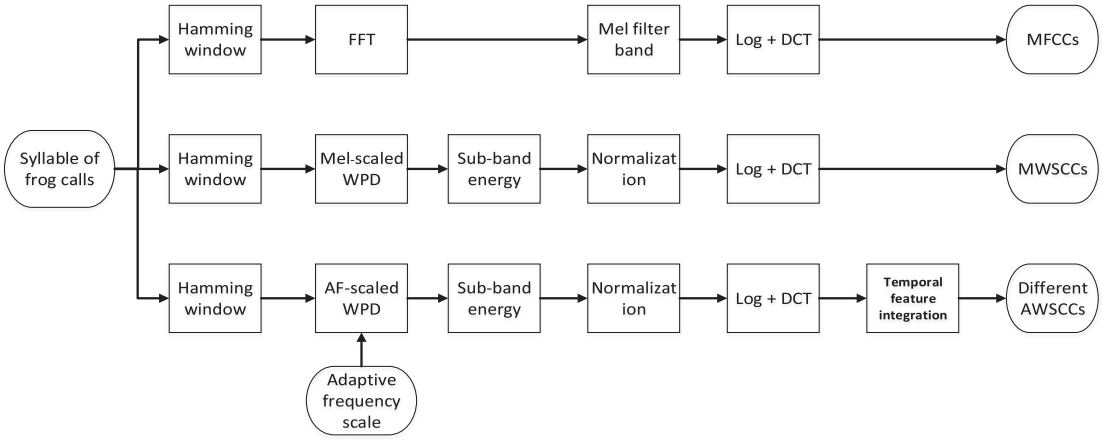


Fig. 6. Description of three feature extraction methods including MFCCs, MWSCCs, and different AWSCCs.

Table 5
Weighted classification accuracy (mean and standard deviation) comparison for five feature sets with two classifiers.

Feature set	Classification accuracy (%)	
	k-NN	SVM
SFs	82.2 ± 11.2	84.2 ± 10.5
MFCCs	90.8 ± 8.6	92.8 ± 11.0
MWSCCs	95.0 ± 7.7	97. ± 5.7
Averaged AWSCCs	98.8 ± 4.2	99.0 ± 3.6
Delta-AWSCCs	99.2 ± 2.1	99.6 ± 1.8

is introduced and summarized as follows. First, the frequency domain boundary is defined based on the dominant frequency, and the power within the boundary is calculated. Then, the power vector is normalized, and the first and last 20% part of the vector is discarded, because of the uncertainty in the start and end of the syllables. Next, the autocorrelation with the length of the vector is calculated. Furthermore, a discrete cosine transform (DCT) is applied to the vector after subtracting the mean, and the position of the highest frequency (P_f) is achieved. Finally, the oscillation rate is defined as

$$O_r = \frac{P_f}{L_{dct}} * r_x * \gamma \tag{5}$$

where P_f is the position of the highest frequency values of the DCT result, L_{dct} is the length for applying DCT to the power vector, and is experientially set as 0.2 s in this study.

2.6. Wavelet packet decomposition

Wavelet packet decomposition (WPD) is a powerful tool for the analysis of non-stationary signals, which includes multiple bases and different basis (Selin et al., 2007). With WPD, an original acoustic signal can be split into two frequency bands such as lower and higher frequency band. Then, both lower and higher frequency bands can be further

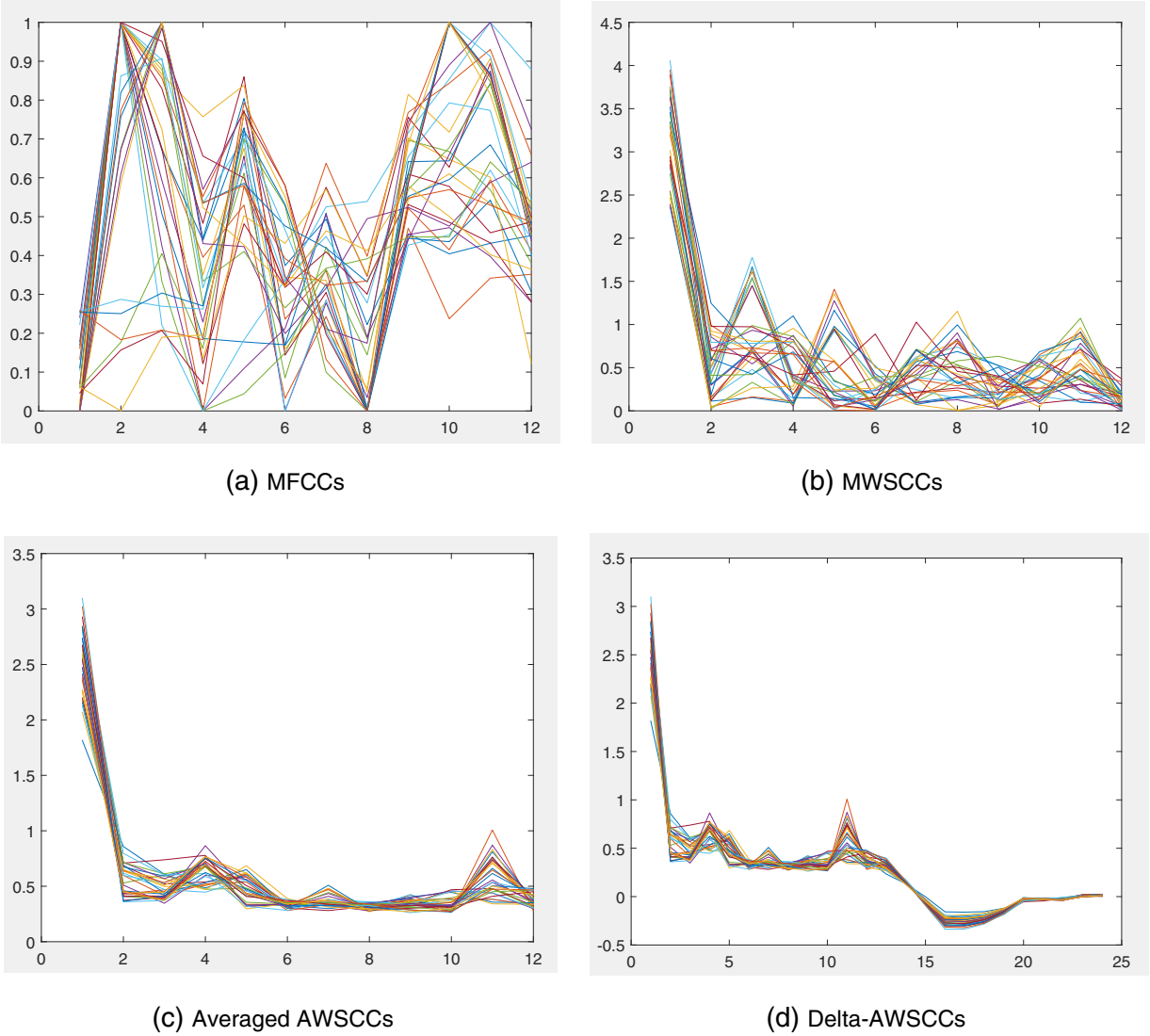


Fig. 7. The feature vectors for 31 syllables of the single species, *Assa darlingtoni*. x-axis is feature index and y-axis is the feature value. Note that the feature vectors for averaged AWSCCs (c) and delta-AWSCCs (d) are more highly correlated than for the other two methods (a) and (b).

continuously decomposed into two sub-bands, which produce a complete wavelet packet tree (Farooq and Datta, 2001). Due to its ability for analyzing a non-stationary signal, WPD has been used to analyze acoustic signals (Selin et al., 2007; Ren et al., 2008). Here, WPD is used to obtain features for frog call classification.

2.7. WPD based on an adaptive frequency scale

To obtain robust features for frog call classification, the frequency scale used for WPD is crucial. In prior work (Litvin and Cohen, 2011; Biswas et al., 2014; Zhang and Li, 2015), different frequency scales have already been proposed for WPD. Bark-scaled WPD was proposed by Litvin and Cohen to separate blind source from a single channel audio source (Litvin and Cohen, 2011). (Biswas et al., 2014) used features based on ERB-scaled (Equivalent rectangular bandwidth) WPD for Hindi consonant recognition. (Zhang and Li, 2015) developed a method based on Mel-scaled WPD for bird sound detection with the SVMs classifier. However, most frequency scales used for WPD are developed for studying speech rather than frogs. Therefore, finding a suitable frequency scale for frogs to perform the WPD is important for obtaining features with strong discriminatory power. In this study, we propose an adaptive frequency scale for WPD for frog calls based on the dominant frequency of frog species to be classified. Specifically, the k-means clustering algorithm is used to cluster the dominant frequency of all syllables. Then, the centroids of the clustering result are used to generate the frequency scale. Here, the value of k for the k-means clustering algorithm is the same as the number of frog species to be classified, the distance function used is *city block* (Melter, 1987).

Based on the obtained frequency scale, an adaptive frequency scaled WPD method is proposed, which is described in Algorithm 1. The wavelet packet tree used for classifying 18 frog species is shown in Fig. 5.

Algorithm 1: Adaptive frequency scale for WPD.

Data: $c_i(i = 1, 2, ..., K)$, f_s, where K is the number of frog species to be classified, c_i is the centroid of the clustering results, $f_s = sr/2$ where sr is the sample rate of the audio recordings, which is 16 kHz here.
Result: Adaptive wavelet packet tree
begin
  Step 1: Sort the centroid $c_i(i = 1, 2, ..., K)$, and calculate the difference between the consecutive vectors of c, sort the difference and save it as $d_j(j = 1, 2, ..., K - 1)$
  Step 2: Calculate the decomposition level L based on the following rule
  $f_s/min(d) \leq 2^{L-1}$
  where L is the minimum integer that satisfies that equation.
  Step 3: Perform the wavelet packet decomposition
  for $l = 1 : L$ **do**
    1. Calculate the frequency resolution of level l
    for $i = 1 : K$ **do**
      1: Put the c_i into the right frequency band
      2: Count the number of c_i in each band (n)
    if $n \geq 2$ **then**
      perform further decomposition to that particular node
    else
      stop decomposition

2.8. Feature extraction based on adaptive frequency scaled WPD

In previous studies (Bedoya et al., 2014; Xie et al., 2015), Mel-frequency cepstral coefficients (MFCCs) have been used for studying bioacoustic data, and used as the baseline for feature comparison in this study. Besides MFCCs, another feature set called Mel-scaled wavelet packet decomposition sub-band cepstral coefficients (MWSCCs) is also included in the comparison experiment (Zhang and Li, 2015), because it shows better performance than MFCCs for bird detection in a complex environment. In this study, we propose a novel feature set named *adaptive frequency scale wavelet packet decomposition sub-band cepstral coefficients* (AWSCCs) for frog call classification. The extraction procedure of AWSCCs is similar to MWSCCs. However, the frequency scale used for our AWSCCs is based on an adaptive frequency scale rather than Mel-scale for MWSCCs. Meanwhile, after performing DCT, temporal feature integration is used for calculating the statistics of feature vectors which generates different AWSCCs (see Fig. 6).

After syllable segmentation, the signal of one syllable is represented as $y(n)$,$n = 1,...,N$, where N is the length of one syllable of frog calls. Based on the $y(n)$, steps for AWSCCs extraction are described as follows:

1) Add Hamming window to the signal $y(n)$.

$$x(n) = w(n)y(n) \tag{6}$$

where $w(L)$ is the Hamming window function and defined as $w(n) = 0.54-0.46\,cos(\frac{2n\pi}{L-1})$, L is the length of Hamming window and set as 128 samples here.

2) Perform wavelet packet decomposition spaced in adaptive frequency scale as described in Section 2.7.

$$WP(i, j) = \sum_{i=1}^{M} x(n)\psi_{(a,b)}(n) \tag{7}$$

where $WP(i,j)$ is the wavelet coefficients of the decomposition, i is the sub-band index, j is the index of wavelet coefficients, $\psi_{(a,b)}(n)$ is the wavelet base function, and we use 'Db 4' experimentally. Here, a and b are the scale and shift parameters, respectively. 'Db 4' represents the Daubechies wavelet transform which has four scaling and wavelet function coefficients.

3) Calculate the total energy of each sub-band.

$$WP_i = \sum_{j=1}^{M_i} [WP(i, j)]^2 \tag{8}$$

where $i = 1,2,...,T$, and T is the total number of sub-band, and $j = 1,2,...,M_i$, M_i is the total number of wavelet coefficients.

4) Normalize the energy of each sub-band.

$$SE_i = \frac{WP_i}{M_i} \tag{9}$$

where $i = 1, 2, \ldots, T$.

Table 6
Classification accuracy of five features for the classification of twenty-four frog species using the SVM classifier. Here, Avg AWSCCs means the averaged AWSCCs.

Code	Classification accuracy (%)				
	SFs	MFCCs	MelCCs	Avg AWSCCs	Delta-AWSCCs
ADI	76.7 ± 15.3	80.0 ± 22.1	83.3 ± 16.7	100.0 ± 0.0	100.0 ± 0.0
CPA	86.7 ± 16.3	100.0 ± 0.0	93.3 ± 13.3	100.0 ± 0.0	100.0 ± 0.0
LCA	93.3 ± 15.3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
LCS	70.0 ± 23.3	63.3 ± 27.7	96.7 ± 10.0	93.3 ± 13.3	96.7 ± 10.0
LFX	91.7 ± 8.3	93.3 ± 8.2	93.3 ± 8.2	100.0 ± 0.0	100.0 ± 0.0
LGA	30.0 ± 45.8	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
LLA	92.7 ± 8.1	98.7 ± 2.7	98.0 ± 4.3	100.0 ± 0.0	100.0 ± 0.0
LNA	78.6 ± 14.6	94.3 ± 9.5	95.7 ± 9.1	100.0 ± 0.0	100.0 ± 0.0
LRA	40.0 ± 30.0	10.0 ± 20.0	100.0 ± 0.0	90.0 ± 20.0	98.2 ± 6.5
LUA	60.0 ± 20.0	100.0 ± 0.0	86.7 ± 22.1	100.0 ± 0.0	100.0 ± 0.0
LVV	100.0 ± 0.0	96.7 ± 10.0	80.0 ± 22.1	93.3 ± 13.3	100.0 ± 0.0
MFS	90.0 ± 15.3	76.7 ± 21.3	90.0 ± 15.3	100.0 ± 0.0	100.0 ± 0.0
MFI	90.0 ± 30.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
PKN	90.0 ± 20.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
PCA	72.5 ± 20.8	77.5 ± 20.8	95.0 ± 10.0	92.5 ± 11.5	100.0 ± 0.0
PRI	45.0 ± 35.0	80.0 ± 33.2	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
RSS	50.0 ± 50.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
ULA	93.3 ± 13.3	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0

Table 7
Paired statistical analysis of the results in Table 4. For the classification accuracy of each frog species, the paired Student t-test was conducted (Tanton, 2005).

Pairs	Test results
Delta-AWSCCs - Avg AWSCCs	$t = 1.95$ (not significant)
Delta-AWSCCs - MWSCCs	$t = 3.41$ (significant at $p < 0.01$, $df = 17$)
Delta-AWSCCs - MFCCs	$t = 2.91$ (significant at $p < 0.01$, $df = 17$)
Delta-AWSCCs - SFs	$t = 5.52$ (significant at $p < 0.001$, $df = 17$)

5) Perform DCT on the logarithm sub-band energy for dimension reduction and obtain the feature AWSCCs.

$$AWSCCs(d) = \sum_{i=1}^{T} logSE_i \cos\left(\frac{d(i-0.5)}{T}\pi\right) \tag{10}$$

where $d = 1, 2, \ldots, d'$, $1 \le d' \le T$, here d' is the dimension of AWSCCs, and set as 12 here. To keep the feature dimension consistency, the dimensions for MFCCs and MWSCCs are also set as 12 in this study, and the detailed steps for extraction can be found in (Bedoya et al., 2014) and (Zhang and Li, 2015).

6) Temporal feature integration
Here, the statistics of all feature vectors over each windowed signal are calculated, which include sum, average, standard deviation, and skewness. With randomly selected five instances for each frog species, the classification accuracy of averaged AWSCCs is higher than other statistics of AWSCCs. Therefore, only averaged AWSCCs are used in the subsequent experiment. To capture the dynamic information of the frog calls, the delta-AWSCCs are also calculated based on the averaged AWSCCs.

2.9. Classification

In this study, the k-nearest neighbour (k-NN) and support vector machine (SVM) classification algorithm are used for frog call classification. The input parameters for each classifier are syllable features (SFs), MFCCs, MWSCCs, and different AWSCCs, and the output is the frog species.

2.9.1. k-nearest neighbours

For the k-NN classifier, an object is classified to the class of the majority of its k-nearest neighbours (Huang et al., 2009). Specifically, frog feature vectors are stored with species labels in the training phase. For the test phase, the distances between an input frog feature vector and all stored vectors are calculated. Then, k closest vectors are used for selecting the most frequent vector as the label. For example, the Euclidean distance between an input feature vector $f_{i,c}$ and one stored feature vector $f_{i,c}$ is calculated as

$$d(i,j) = \sqrt{\sum_{c=1}^{n} \left(f_{i,c} - f_{j,c}\right)^2} \tag{11}$$

where i and j are indices of the feature vector, n means the dimension of the feature vector. Next, k nearest neighbours of the feature vector i are selected based on the Euclidean distance for selecting the most frequent vector as the label. If the following equation is satisfied

$$\frac{1}{k_1}\sum_{j \in s_1} d(i, j(s_1)) \le \frac{1}{k_2}\sum_{j \in s_2} s(i, j(s_2)) \tag{12}$$

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 feature vector i will be classified as frog species s_2.

2.9.2. Support vector machines

Due to the high accuracy and superior generalization properties, support vector machines have been widely used for classifying animal sounds (Huang et al., 2009; Acevedo et al., 2009). In this study, the feature set obtained is first selected as training data. Then, the pairs $(v_l^n, L_l^n), l = 1, 2, \ldots, C_l$ are constructed using the selected training data, where C_l is the number of frog instance in the training data, v_l^n is the feature vector obtained from the l-th frog instance in the training data,

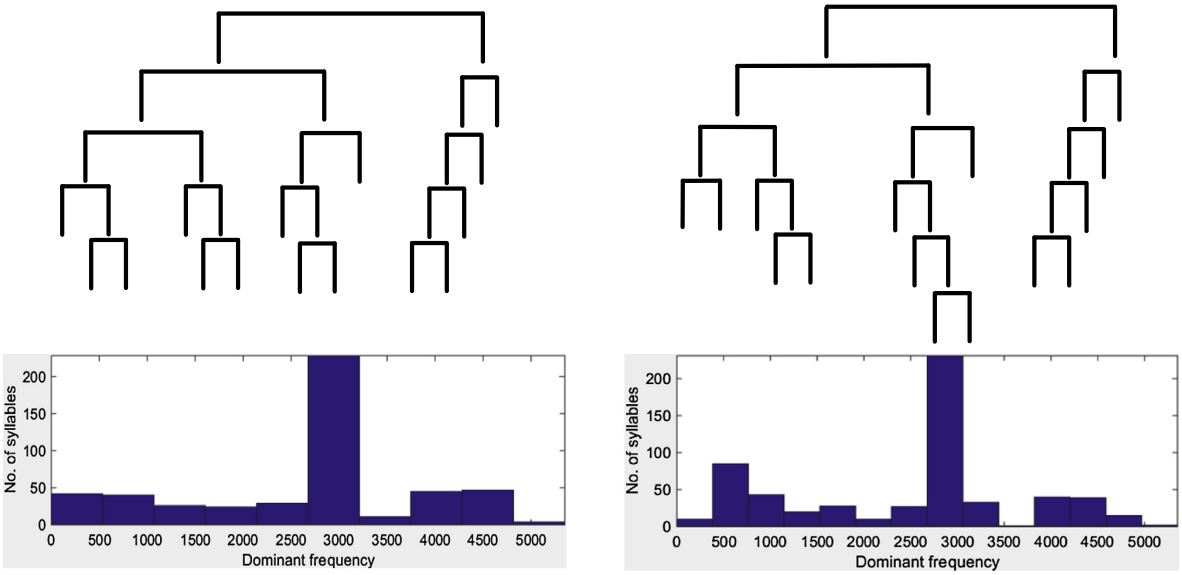


Fig. 8. Wavelet packet tree based on adaptive frequency scale for classifying 10 and 15 frog species.

L_l^n is the frog species label. Furthermore, the decision function for the classification problem based on SVM (Cortes and Vapnik, 1995) is defined by the training data as follows:

$$f(v) = sgn\left(\sum_{sv} \alpha_l^n L_l^n K(v, v_l^n) + b_l^n\right) \quad (13)$$

where $K(.,.)$ is the kernel function, α_l^n is the Lagrange multiplier, and b_l^n is the constant value.

3. Experiment result and discussion

Several experiments are described for evaluating our proposed frog call classification system. First, the parameter tuning is discussed based on the reference data set. Then, the comparisons between all proposed features are studied. Finally, the classification results under different SNR are described.

3.1. Parameter tuning

Five modules for parameter tuning are syllable segmentation, spectral peak track, feature extraction, and classification (Fig. 1).

For syllable segmentation, the window size and overlap are 512 samples and 25%, however, the intensity threshold is 10 dB and 5 dB for the commercial recordings and the JCU recordings respectively.

In the spectral peak track determination, there are seven parameters (see in Table 3). The parameter settings are shows in Table 4.

With a random parameter setting start, an iterative loop is performed for a fixed range of each parameter base on Table 1 to optimise those parameters.

For feature extraction, the window size and overlap are the same for MFCCs, MWSCCs, and AWSCCs using Hamming window, which are 128 samples and 90%, respectively. The dimensions of MFCCs, MWSCCs and AWSCCs are 12. For SFs and delta-AWSCCs, the dimensions are 3 and 24, respectively.

Following prior work (Huang et al., 2009; Han et al., 2011; Xie et al., 2015), the distance function used for k-NN is the Euclidean distance, and k is set as 3. As for the SVM classifier, the Gaussian kernel is used. Parameters α and v are selected independently for each feature set by grid-search using cross validation (Hsu et al., 2003).

3.2. Feature evaluation

All experiments are carried out in Matlab R2014b. Performance statistics are estimated with ten-fold cross validation. Totally, five feature sets including SFs, MFCCs, MWSCCs, and averaged AWSCCs, and delta-AWSCCs, are adopted to two classifiers, which are the k-NN and SVM classifiers. Both k-NN and SVM classifiers are run ten times for evaluating the feature robustness. Due to the non-uniform distribution of the number of syllables for different frog species in the commercial recordings, a weighted classification accuracy is define as follows

$$weighted\ \ Acc = \sum_{i=1}^{N} Acc(i) * \frac{n_i}{N} \quad (14)$$

where n_i is the number of syllables for frog species i, N is the number of syllables for all frog species, Acc is the classification accuracy for that particular frog species.

3.3. Comparison between different feature sets

The classification accuracy comparison for 18 frog species using five feature sets and two classifiers are shown in Table 5.

In this experiment, the best classification accuracy is 99.6%, which is achieved by the delta-AWSCCs with the SVM classifier. Compared with the average AWSCCs, the delta-AWSCCs are slightly improved. One may conjecture that the delta-AWSCCs can capture the dynamic information of the frog calls. For MWSCCs, the averaged classification accuracy of both classifiers is about 2% lower than averaged AWSCCs and delta-AWSCCs with 96.3%. The improvement shows that our proposed adaptive frequency scale can capture more information of frog calls than Mel-scale (Fig. 7).

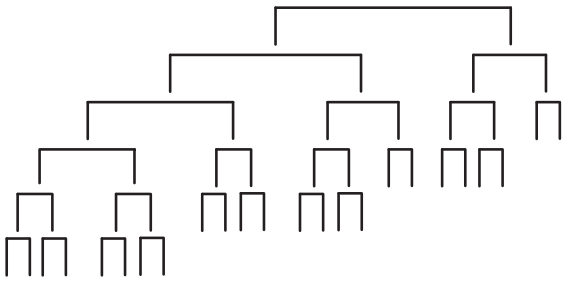


Fig. 9. Mel-scaled wavelet packet tree for frog call classification.

As for SFs and MFCCs, the averaged classification accuracy is much lower than AWSCCs, which is 83.2% and 91.8%, respectively. To explore the reason for the improvement of our proposed feature, the frog call classification accuracy of all frog species is shown in Table 6. However, only the features that use the SVM classifier is shown, because averaged accuracy of the k-NN classifier (93.2%) is lower than the SVM classifier (94.64%).

Table 6 lists the classification accuracy of all 18 frog species with five features. It can be seen from the table that delta-AWSCCs have an accuracy greater than 95% for all frog species. Compared with averaged AWSCCs, the classification accuracy of *Pseudophryne coriacea* (PCA) and *Litoria verreauxii verreauxii* (LVV) are improved to 100%, it might be that the delta-AWSCCs include the dynamic information of frog calls. For *Litoria revelata* (LRA), both the classification accuracy of averaged AWSCCs and delta-AWSCCs are lower than 100%, it is because the dominant frequency is quite similar with multiple frog species including *Assa darlingtoni* (ADI), *Litoria nasuta* (LNA) and *Litoria verreauxii verreauxii* (LVV). However, the classification of *Litoria revelata* (LRA) is 100% using Mel-scale based techniques, because the Mel-scale has a better frequency resolution for *Litoria chloris* (LCS) within its dominant frequency range. In Table 8, the classification accuracy of SFs and MFCCs is lower than other three features, which is only 84.2% and 92.8%, respectively.

The statistical significance of the results is shown in Table 7. The classification accuracy of average AWSCCs is not significantly lower than the delta-AWSCCs. However, the classification accuracy of MWSCCs, MFCCs and SFs is significantly lower than delta-AWSCCs.

Since our wavelet packet tree for feature extraction is obtained based on the frog species to be classified, two more experiments are used for further evaluation. The first experiment is to classify first ten frog species (No.1–10); the second is to classify the first fourteen frog species (No.1–14) (see Table 1). The wavelet packet tree for classifying ten and fourteen frog species is shown in Fig. 8, which is different from the tree for classifying eighteen frog species. However, the Mel-scaled wavelet packet tree is the same for all experiments (see Fig. 9). The classification results are shown in Table 8. Since the classification accuracy

Table 8
Classification accuracy (%) for classifying different number of frog species with four feature sets.

Features	SFs	MFCCs	MWSCCs	Averaged AWSCCs
frog species	84.2 ± 10.5	92.8 ± 11.0	97.6 ± 5.7	99.0 ± 4.6
frog species	89.6 ± 9.7	94.4 ± 8.5	99.2 ± 2.6	100.0 ± 0.0
frog species	94.6 ± 8.7	95.8 ± 8.6	100.0 ± 0.0	100.0 ± 0.0

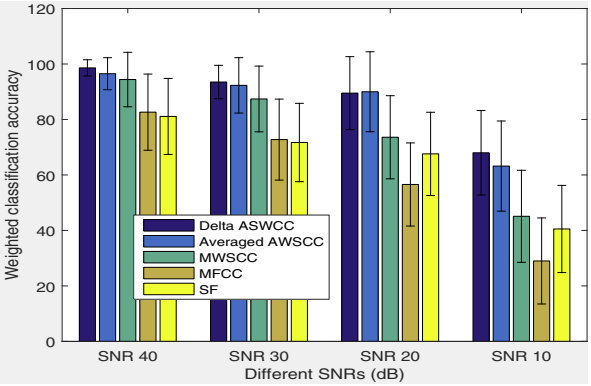


Fig. 10. Sensitivity of five features for different levels of noise contamination.

with averaged AWSCCs is very high for classifying 10 and 14 frog species, the delta-AWSCCs is not included in this experiment. Table 8 shows that averaged AWSCCs can achieve the highest classification accuracy for classifying different number of frog species. Since the averaged AWSCCs is adaptively extracted based on the data, more frog species do not cause a large decrease in the classification accuracy.

3.4. Comparison under different SNRs

To further evaluate the robustness of the proposed feature, a Gaussian noise signal, with SNR of 40 dB, 30 dB, 20 dB, and 10 dB, is added to the original signal. The noise is added after syllable segmentation, because this study focuses on the development of novel features for classification rather than the segmentation method. The classification accuracy with five features under different SNRs is shown in Fig. 10. Compared with MFCCs and MWSCCs, SFs has a stronger anti-noise performance, because the dominant frequency of SFs has a small variation under low SNR. Correspondingly, the adaptive frequency scale also has a small variation, because it is generated by means of applying the k-means clustering algorithm to the dominant frequency. Therefore, our proposed feature has a stronger anti-noise performance than other cepstral features (MFCCs and MWSCCs).

3.5. Proposed feature evaluation using the JCU recordings

Table 9 shows the classification accuracy comparison using our proposed feature to classify 8 frog species obtained from the JCU recordings. Since calls of some frog species in the JCU recordings do not have oscillation structure, SFs are not included for the comparison. Compared with other referred features, our proposed feature also achieves the best classification performance. Since the JCU recordings often have multiple calls from different frog species, spectral peak track occasionally can not capture the specific frog species (labelled species for that syllable) but other frog species to be classified, however, applying k-mean clustering to the dominant frequency calculated from the spectral peak track can reduce this deviation. Therefore, the frequency scale used for the WPD can be accurately achieved which still leads to a high classification accuracy with the proposed feature.

Table 9
Classification accuracy using the JCU recordings.

Feature set	Classification accuracy (%)	
	k-NN	SVM
MFCCs	67.5 ± 13.2	70.8 ± 14.1
MWSCCs	90.4 ± 9.2	91.6 ± 8.7
Averaged AWSCCs	94.1 ± 6.3	94.5 ± 5.8
Delta-AWSCCs	97.0 ± 5.2	97.4 ± 5.4

4. Conclusion and future work

In this study, a novel feature extraction method for frog call classification is developed using the adaptive frequency scaled wavelet packet decomposition. With segmented syllables, spectral peak track is first extracted from each syllable. Then, track duration, dominant frequency, and oscillation rate are calculated based on each track. Next, a k-means clustering algorithm is applied to the dominant frequency, which generates the frequency scale for WPD. Finally, a new feature set, AWSCCs, is calculated. Since our feature extraction method is developed based on the data itself, the wavelet packet tree differs according to the frog species to be classified. Compared with the Mel-scaled WPD tree, the proposed adaptive wavelet packet tree can better fit the dominant frequency distribution of the frog species to be classified. With the proposed frequency scale, the call character of those frog species to be classified can be enhanced, however, the background noise and calls from other animals will be suppressed. Therefore, our proposed feature sets can achieve a higher accuracy for the classification of frog calls than others. Meanwhile, since the frequency scale is calculated based on the dominant frequency of those frog species to be classified, our proposed wavelet tree structure is more accurate and efficiency in classifying the frog calls when compared with Mel-scale (Figs. 8 and 9).

As for the feature extraction algorithm, it is designed for classifying frog calls. For frog calls, the typical structure in a spectrogram is frequency contour (named spectral peak track in this study) that are within a given frequency range starting at a given time (Mellinger et al., 2011). For other organisms that have similar frequency contour structures such as the whistles of dolphins, chirps of birds (Chen and Maher, 2006), spectral peak tracks can also be extracted from the spectrograms of their calls. Based on those spectral peak tracks, dominant frequency can be calculated. For the subsequent analysis, we can calculate the features using the same process as described in this study. For those organisms without clear frequency contour structure, this proposed method can also be used by enhancing the frequency contour structure, which can be realized by applying a small window size and a large window overlap to the recording waveform.

For future work, the oscillation rate is calculated based on the spectrogram, which is generated by applying STFT to the waveform. However, when the temporal gap is smaller than the window size used for STFT, the oscillation structure will disappear. Therefore, finding new techniques for translating the 1-D signal to 2-D signal is our future direction. Since the frequency scale is generated based on the dominant frequency, this technique can be applied to other organisms that have clear frequency contour structure. Modifying this algorithm to those organisms without a clear frequency contour structure needs to be solved. We also plan to include additional experiments that test a wider variety of audio data from different geographical and environment conditions. Other animal calls such as birds, insects, and whales can also be studied. Furthermore, we will explore the idea of developing new features based on the data itself.

Acknowledgement

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

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J., Aide, T.M., 2009. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecol. Inf. 4, 206–214.

Bedoya, C., Isaza, C., Daza, J.M., López, J.D., 2014. Automatic recognition of anuran species based on syllable identification. Ecol. Inf. 24, 200–209.

Biswas, A., Sahu, P., Chandra, M., 2014. Admissible wavelet packet features based on human inner ear frequency response for hindi consonant recognition. Comput. Electr. Eng. 40, 1111–1122.

Chen, Z., Maher, R.C., 2006. Semi-automatic classification of bird vocalizations using spectral peak tracks. J. Acoust. Soc. Am. 120, 2974–2984.

Chen, W.P., Chen, S.S., Lin, C.C., Chen, Y.Z., Lin, W.C., 2012. Automatic recognition of frog calls using a multi-stage average spectrum. Comput. Math. Appl. 64, 1270–1281.

Clauzel, C., Bannwarth, C., Foltete, J.C., 2015. Integrating regional-scale connectivity in habitat restoration: An application for amphibian conservation in eastern france . J. Nat. Conserv. 23, 98–107. http://dx.doi.org/10.1016/j.jnc.2014.07.001.

Colonna, J., Ribas, A., dos Santos, E., Nakamura, E., 2012. Feature subset selection for automatically classifying anuran calls using sensor networks. Neural Networks (IJCNN), The 2012 International Joint Conference on, pp. 1–8 http://dx.doi.org/10.1109/IJCNN.2012.6252794.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Farooq, O., Datta, S., 2001. Mel filter-like admissible wavelet packet structure for speech recognition. IEEE Signal Process Lett. 8, 196–198.

Gage, S.H., Axel, A.C., 2014. Visualization of temporal change in soundscape power of a michigan lake habitat over a 4-year period. Ecol. Inf. 21, 100–109.

Garcia, R.A., Cabeza, M., Rahbek, C., Araújo, M.B., 2014. Multiple dimensions of climate change and their implications for biodiversity. Science 344, 1247579.

Gingras, B., Fitch, W.T., 2013. A three-parameter model for classifying anurans into four genera based on advertisement calls. J. Acoust. Soc. Am. 133, 547–559.

Han, N.C., Muniandy, S.V., Dayou, J., 2011. Acoustic classification of australian anurans based on hybrid spectral-entropy approach. Appl. Acoust. 72, 639–645.

Harma, A., 2003. Automatic identification of bird species based on sinusoidal modeling of syllables. Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on, IEEE., pp. V–545

Heller, J.R., Pinezich, J.D., 2008. Automatic recognition of harmonic bird sounds using a frequency track extraction algorithm. J. Acoust. Soc. Am. 124, 1830–1837. http://dx.doi.org/10.1121/1.2950085.

Hsu, C.W.W., Chang, C.C., Lin, C.J., et al., 2003. A practical guide to support vector classification.

Huang, C.J., Yang, Y.J., Yang, D.X., Chen, Y.J., 2009. Frog classification using machine learning techniques. Expert Syst. Appl. 36, 3737–3743.

Jancovic, P., Kokuer, M., 2015. Acoustic recognition of multiple bird species based on penalized maximum likelihood. IEEE Signal Process Lett. 22, 1585–1589. http://dx.doi.org/10.1109/LSP.2015.2409173.

Litvin, Y., Cohen, I., 2011. Single-channel source separation of audio signals using bark scale wavelet packet decomposition. J. Signal Proc. Sys. 65, 339–350.

Mellinger, D.K., Martin, S.W., Morrissey, R.P., Thomas, L., Yosco, J.J., 2011. A method for detecting whistles, moans, and other frequency contour sounds. J. Acous. Soc. Am. 129, 4055–4061.

Melter, R.A., 1987. Some characterizations of city block distance. Pattern Recogn. Lett. 6, 235–240.

Ren, Y., Johnson, M.T., Tao, J., 2008. Perceptually motivated wavelet packet transform for bioacoustic signal enhancement. J. Acoust. Soc. Am. 124, 316–327.

Roch, M.A., Brandes, T.S., Patel, B., Barkley, Y., Baumann-Pickering, S., Soldevilla, M.S., 2011. Automated extraction of odontocete whistle contours. J. Acoust. Soc. Am. 130, 2212–2223.

Sahidullah, M., Saha, G., 2012. Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. Speech Comm. 54, 543–565. http://dx.doi.org/10.1016/j.specom.2011.11.004.

Selin, A., Turunen, J., Tanttu, J.T., 2007. Wavelets in recognition of bird sounds. EURASIP J. Appl. Signal Proc. 2007, 141-141.

Shine, R., 2014. A review of ecological interactions between native frogs and invasive cane toads in australia. Austral Ecol. 39, 1–16.

Stewart, D., 1999. Australian frog calls: subtropical east. Audio CD . URL: http://www.naturesound.com.au/cd_frogsSE.htm.

Tanton, J.S., 2005. Encyclopedia of mathematics. Facts On File.

Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P., 2012. A toolbox for animal call recognition. Bioacoustics 21, 107–125.

Wimmer, J., Towsey, M., Planitz, B., Roe, P., Williamson, I., 2010. Scaling acoustic data analysis through collaboration and automation. e-Science (e-Science), 2010 IEEE Sixth International Conference on, pp. 308–315 http://dx.doi.org/10.1109/eScience.2010.17.

Wimmer, J., Towsey, M., Planitz, B., Williamson, I., Roe, P., 2013. Analysing environmental acoustic data through collaboration and automation. Futur. Gener. Comput. Syst. 29, 560–568.

Xie, J., 2016. Widelife acoustic. http://www.wildlifeacoustics.com/products/song-meter-sm2-birds.

Xie, J., Towsey, M., Truskinger, A., Eichinski, P., Zhang, J., Roe, P., 2015. Acoustic classification of australian anurans using syllable features. 2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (IEEE ISSNIP 2015), Singapore, Singapore.

Yen, G.G., Fu, Q., 2002. Automatic frog call monitoring system: a machine learning approach. AeroSense 2002, International Society for Optics and Photonics., pp. 188–199

Yen, G.G., Lin, K.C., 2000. Wavelet packet feature extraction for vibration monitoring. IEEE Trans. Ind. Electron. 47, 650–667.

Zhang, X., Li, Y., 2015. Adaptive energy detection for bird sound detection in complex environments. Neurocomputing 155, 108–116.

Zhang, J., Huang, K., Cottman-Fields, M., Truskinger, A., Roe, P., Duan, S., Dong, X., Towsey, M., Wimmer, J., 2013. Managing and analysing big audio data for environmental monitoring. Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on, pp. 997–1004.

Chapter 6

Multiple-instance multiple-label learning for the classification of frog calls with acoustic event detection

---

6.1 Introduction

This chapter presents a method for the classification of simultaneously vocalising frog species in low SNR recordings. In chapter 4 and 5, frog call classification is solved using a SISL framework, which cannot reflect the nature of the low SNR recordings. Most low SNR recordings often consist of multiple simultaneously animal vocal activities including frogs, birds, crickets and so on. This character of low SNR recordings makes the multiple-instance multiple-label (MIML) learning a suitable classification framework for addressing. To be specific, individual frog syllables in one audio clip is regarded as *multiple instance*, and the frog species included in that audio clip denotes *multiple labels*. The key part of this MIML classification framework for frog calls is to detect individual syllables. After syllables detection, standard acoustic features and MIML classifiers can then be used to perform the MIML classification.

To evaluate our proposed classification framework, a representative sample of 342 10-seconds recordings was exported from the database and split into testing and training sets. The performance is evaluated based on the MIML learning measures. Experimental results demonstrate the MIML classification framework can be adopt to classify multiple simultaneously vocalising frog species in low SNR recordings.

6.2 Conference paper - Multiple-instance multiple-label learning for the classification of frog calls with acoustic event detection

Multiple-Instance Multiple-Label Learning for the Classification of Frog Calls With Acoustic Event Detection

Jie Xie, Michael Towsey, Liang Zhang, Kiyomi Yasumiba, Lin Schwarzkopf, Jinglan Zhang, and Paul Roe

Electrical Engineering and Computer Science School, Queensland University of Technology, Brisbane, Australia
xiej8734@gmail.com
https://www.ecosounds.org/

Abstract. Frog call classification has received increasing attention due to its importance for ecosystem. Traditionally, the classification of frog calls is solved by means of the single-instance single-label classification classifier. However, since different frog species tend to call simultaneously, classifying frog calls becomes a multiple-instance multiple-label learning problem. In this paper, we propose a novel method for the classification of frog species using multiple-instance multiple-label (MIML) classifiers. To be specific, continuous recordings are first segmented into audio clips (10 seconds). For each audio clip, acoustic event detection is used to segment frog syllables. Then, three feature sets are extracted from each syllable: mask descriptor, profile statistics, and the combination of mask descriptor and profile statistics. Next, a bag generator is applied to those extracted features. Finally, three MIML classifiers, MIML-SVM, MIML-RBF, and MIML-kNN, are employed for tagging each audio clip with different frog species. Experimental results show that our proposed method can achieve high accuracy (81.8% true positive/negatives) for frog call classification.

Keywords: Frog call classification, acoustic event detection, multiple-instance multiple-label learning

1 Introduction

Recently, human activity and climate change put a negative effect on frog biodiversity, which makes frog monitoring become ever more important. Compared with the traditional monitoring method such as field observation, acoustic sensors have greatly extended acoustic monitoring into larger spatio-temporal scales [1]. Correspondingly, large volumes of acoustic data are generated, which makes it essential to develop automatic methods.

Several papers have already described automated methods for the classification of frog calls. Han et.al combined spectral centroid, Shannon entropy, Renyi entropy for frog call recognition with a k-nearest neighbour classifier [2]. Gingras

et al. proposed a method based on mean value for dominant frequency, coefficient of variation of root-mean square energy, and spectral flux for anuran classification [3]. Bedoya et al. used Mel-frequency cepstral coefficients (MFCCs) for the recognition of anuran species with a fuzzy classifier [4]. Xie et al proposed a method based on track duration, dominant frequency, oscillation rate, frequency modulation and energy modulation to do frog call [5]. All those previous methods achieve a high accuracy rate in recognition and classification, but recordings used in those papers are assumed that there is only a single frog species present in each recording.

Unfortunately, all the recordings used in this study are low signal to noise ratio and contain many overlapping animal vocal activities including frogs, birds, crickets and so on. To solve this problem, the multiple-instance multiple-label classifier for supervised classification is formulated [6]. In the previous study, Briggs et al has already introduced the MIML classifiers for acoustic classification of multiple simultaneous bird species [7]. In their method, a supervise learning classifier was employed for segmenting acoustic events, which required lots of annotations.

In this study, we introduced the MIML algorithm for frog call classification. Rather than using a supervised learning method for syllable segmentation, acoustic event detection is first employed to separate frog syllables. Then, three feature sets, mask descriptor, profile statistics, and the combination of mask descriptor and profile statistics, are calculated from each syllable. After applying a bag generator to those extracted feature sets, three classifiers, MIML-SVM [6], MIML-RBF [8], and MIML-kNN [9], are lastly used for the recognition of multiple simultaneous frog species. Experimental results show that our proposed method can achieve high classification accuracy.

2 Materials and methods

2.1 Materials

Digital recordings in this study were obtained with a battery-powered, weatherproof Song Meter (SM2) box. Recordings were two-channel, sampled at 22.05 kHz and saved in WAC4 format. Here, a representative sample of 342 10-s recordings was selected to train and evaluate our proposed algorithm for predicting which frog species are present in a recording. All those examples were collected between 02/2014 to 03/2014, because it is the frog breeding season with high calling activity. All the species that are present in each 10-s recording were manually labelled by an ecologist who studies frog calls. There are totally eight frog species in the recordings: Canetoad (CAD) (F_0=560 Hz), Cyclorana novaehollandiae (CNE) (F_0=610 Hz), Limnodynastes terraereginae (LTE) (F_0=610 Hz), Litoria fallax (LFX) (F_0=4000 Hz), Litoria nasuta (LNA) (F_0=2800 Hz), Litoria rothii (LRI) (F_0=1800 Hz), Litoria rubella (LRA) (F_0=2300 Hz), and Uperolela mimula (UMA) (F_0=2400 Hz). Here, F_0 is the mean dominant frequency for each frog species. Each recording contains between 1 and 5 species. Following

the prior work [7], we assume that recordings without any frog calls can be detected by acosutic event detection.

2.2 Signal processing

All the recordings were re-sampled at 16 kHz and mixed to mono. A spectrogram was then generated by applying short-time Fourier transform to each recording. Specifically, each recording was divided into frames of 512 samples with 50% frame overlap. A fast Fourier transform was then performed on each frame with a Hamming window, which yielded amplitude values for 256 frequency bins, each spanning 31.25 Hz. The final decibel values (S) were generated using $S_{tf} = 20 * log_{10}(A_{tf})$, where A is the amplitude value, $t = 0, ..., T-1$ and $f = 0, ..., F-1$ represent frequency and time index, T and F are 256 frequency bins and 625 frames, respectively.

2.3 Acoustic event detection for syllable segmentation

Acoustic event detection (AED) aims to detect specified acoustic event in an audio stream. In this study, we use AED to segment frog syllables. Since all the recordings are collected from the field, there are much overlapping vocal activities. Traditional methods for audio segmentation are based on time domain information [10, 11], which cannot address those recordings. Here, we modified the AED method developed by Towsey et al [12] to segment recordings with overlapping activities. The detail of our AED method is described as follows:
Step 1: Wiener filter
To de-noise and smooth the spectrogram, a 2-D Wiener filter is applied to the spectrogram image over a 5×5 time-frequency grid, where the filter size is selected after considering the trade-off between removing the background graininess and blurring the acoustic events.

$$\hat{S_{tf}} = \mu + \frac{(\sigma^2 - \nu^2)}{\sigma^2}(S_{tf} - \nu) \tag{1}$$

where μ and σ^2 are local mean and variance, respectively. ν^2 is the noise variance estimated by averaging all local variances.
Step 2: Spectral subtraction
After Wiener filter, the graininess has been removed. However, some noises such as wind, insect, motor engine that cover the whole recording cannot be removed. Here, a modified spectral subtraction is used for dealing with those noise [13].
Step 3: Adaptive thresholding
After noise reduction, the next step is to convert the noise reduced spectrogram $\hat{S_{tf}^{'}}$ into the binary spectrogram S_{tf}^{b} for events detection. Different from the hard threshold in Towseys work, an adaptive thresholding method named *Otsu thresholding* is used to convert the smoothed spectrogram into binary spectrogram. Otsus method assumes that the spectrogram is composed of two classes: acoustic events and background noise. An optimal threshold value is used for

Algorithm 1: Modified Spectral Subtraction

Data: $\hat{S_{tf}}$, spectrogram after Wiener filtering.
Result: $\hat{S_{tf}^{'}} = \hat{S_{tf}}$, noise reduced spectrogram.
begin
  Construct an array of the modal noise values for all frequency bins;
  for $f \in F$ **do**
    1. calculate the histogram of the intensity value over each frequency bin
    2. smooth the histogram array with a moving average window of size 7
    3. regard the modal noise intensity at the position of maximal bin in the left-side of the histogram
  Smooth the array with a moving average filter with window of size 5;
  for $f \in F$ **do**
    1. subtract the modal noise intensity
    2. truncated negative decibel values to zero

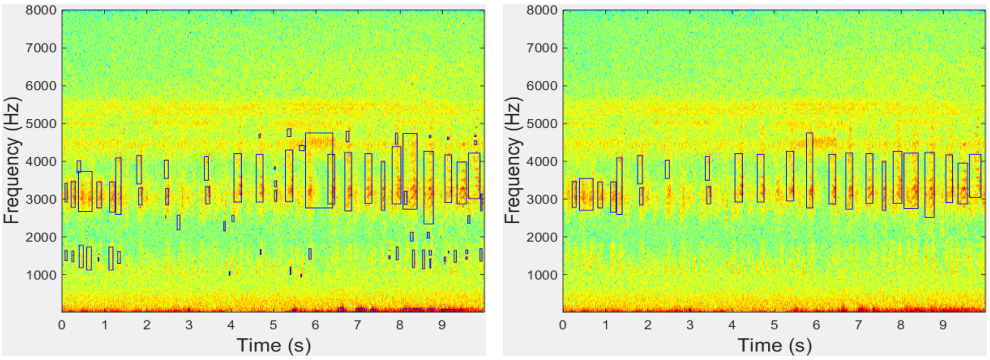


Fig. 1. Acoustic event detection results before (Left) and after (Right) event filtering based on dominant frequency. Here, blue rectangle means the time and frequency boundary of each detected event.

the decision. After thresholding, each group of contiguous positive pixels will be regarded as a candidate event.

Step 4: Events filtering using dominant frequency and event area

After aforementioned process, not all detected events are correspond to frog vocalizations. To further remove those events that are from the listed frog species in section 2.1, dominant frequency (F_0) and area within the event boundary (Ar) are used for filtering.

Step 5: Region growing

Region growing algorithm is utilized to obtain the contour of the particular acoustic event [14]. To get the accuracy boundary of each acoustic event and improve the discrimination of extracted features, a 2-D region growing algorithm is applied for obtaining the accuracy event shape within each segmented event. First, a maximal intensity value within each segmented event is selected as the seed point. Then, if the difference between the neighbourhood pixels and the seed(s) is smaller than the threshold, the neighbourhood pixels will be located

and assigned to the output image. Next, the new added pixels are used as seeds for further processing until all the pixels that satisfy the criteria are added to the output image. The final results after region growing are shown in Fig. 2. Here, the threshold value is empirically set as 5 dB.

Algorithm 2: Event filtering based on dominant frequency and event area

Data: S^b_{tf}, spectrogram; $t_s(n)$, $t_e(n)$, $f_l(n)$, $f_h(n)$, location of each acoustic event n; $F_0(i)$, dominant frequency of frog species i.
Result: $\tilde{S_{tf}}$, spectrogram after events filtering.
begin
  Calculate the area of each acoustic event n.
  $Area(n) = (t_e(n) - t_s(n)) * (f_h(n) - f_l(n))$
  for $n \in N_{e1}$ **do**
    if $Ar(n) \geq Ar_l$ **then**
      split event n into small events
  where Ar_l is set as 3000 pixels.
  Filter events using dominant frequency $f_d(n) = \sum_{t=t_s(n)}^{t_e(n)} F(t)/t_e(n) - t_s(n)$
  where $F(t)$ is the peak frequency of each frame within the event area
  for $n \in N_{e2}$ **do**
    for $i \in I$ **do**
      if $f_d(n) \geq F_0(i) + \theta$; $f_d(n) \leq F_0(i) - \theta$ **then**
        $f_d(n) = 0$;
  where θ is frequency range and set as 300 Hz.
  Remove small acoustic events except frequency band between θ_l and θ_h
  for $n \in N_{e2}$ **do**
    if $Ar(n) \leq Ar_s$ **then**
      remove event n
  where Ar_s is set at 300 pixels, θ_l and θ_h are set as 300 Hz and 800 Hz, respectively. Because the area of LTE is smaller than Ar_s.

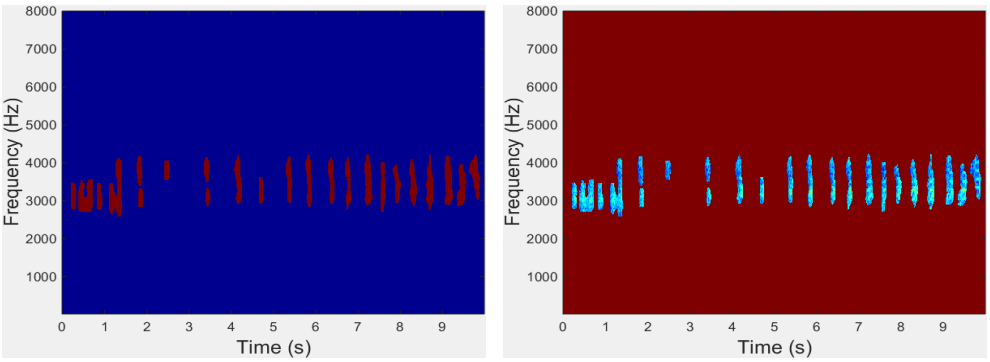


Fig. 2. Acoustic event detection results after region growing. Left: binary segmentation results; Right: segmented frog syllables.

2.4 Feature extraction

Based on acoustic event detection results, two feature sets are first calculated to describe each event (syllable): mask descriptor and profile statistic [7]. Here, we exclude histogram of orientation from our feature sets, because the previous study has already demonstrated its poor classification performance [7]. For mask descriptor, it is used to describe the syllable shape including minimum frequency, maximum frequency, bandwidth, duration, area, perimeter, non-compactness, rectangularity. For profile statistics, there are time-Gini, frequency-Gini, frequency-mean, frequency-variance, frequency-skewness, frequency-kurtosis, frequency-max, time-max, mask-mean, and mask standard deviation. The third feature set consists of all features.

Table 1. Accuracy measure for MIML classifiers with different feature sets. Here, ↓ indicates the smaller the better, while ↑ indicates the bigger the better.

Feature	Algorithm	Hamming loss ↓	Rank loss ↓	One-error ↓	Coverage ↓	Micro-AUC ↑
MD	MIML-SVM	0.253	0.186	0.308	3.147	0.745
MD	MIML-kNN	0.205	0.153	0.298	2.647	0.771
MD	MIML-RBF	**0.182**	**0.132**	**0.223**	**2.352**	**0.828**
PS	MIML-SVM	0.239	0.208	0.323	3.544	0.728
PS	MIML-kNN	0.211	0.153	0.298	2.647	0.777
PS	MIML-RBF	0.186	0.161	0.338	3.161	0.746
AF (MD+PS)	MIML-SVM	0.261	0.199	0.279	3.588	0.761
AF (MD+PS)	MIML-kNN	0.205	0.160	0.264	2.735	0.787
AF (MD+PS)	MIML-RBF	0.191	0.142	0.220	2.632	0.821

3 Multiple-instance multiple-label classifiers

After feature extraction, three MIML algorithms are evaluated for the classification of multiple simultaneous frog calls: MIML-SVM, MIML-RBF, and MIML-kNN. With some form of event-level distance measure, the MIML problem has been reduced to a single-instance multiple-label problem by associating each event with a event-level feature [7]. Here, the maximal and average Hausdorff distances between two syllables are used by MIML-SVM and MIML-RBF, separately. For MIML-kNN, the nearest neighbour is used to assign syllable-level features.

4 Experiment results

4.1 Parameter tuning

There are three modules whose parameters need to be discussed: signal processing, acoustic event detection, and classification. For signal processing, the window size and overlap are 512 samples and 50%, respectively. During the process of acoustic event detection, four thresholds for event filtering need to be

determined, which are small and large area threshold, and frequency boundary for events filtering. All those thresholds were determined empirically by applying various combinations of thresholds to a small number of randomly selected 10s clips. For MIML-SVM classifiers, the parameters used are (C, γ, r) and set as (0.1, 0.6, 0.2) experimentally. For MIML-RBF, the parameters are (r, μ) and set as (0.1,0.6). For MIML-kNN, the number of references (k) and citers ($k^{'}$) are 10 and 20, respectively.

4.2 Classification

In this study, all the algorithms were programmed in Matlab 2014b. Each MIML algorithm is evaluated with five-fold cross-validation on the collection of 342 species-labelled recordings. Five measures including Hamming loss, rank loss, one-error, coverage, and micro-AUC are used to characterize the accuracy of each algorithm [15] [16]. The definition of each measure can be found in [7], the positive/negatives is defined as 1−Hamming loss and it is 0.818 for MIML-RBF with MD. Mask descriptor (MD) and profile statistical (PS), and all features (AF) are put into the three classifiers, respectively. The performance of each MIML classifier is shown in Table 1. Here, the best classification accuracy is achieved by MIML-RBF using MD. For each classifier, the classification accuracy of MD is higher than PS and AF, which shows that the event shape have higher discrimination power than the event content. To give a concrete view of predictions, the results of 5 randomly selected recordings using MIML-RBF are shown in Table 2. Recordings of No.1 and No.3 are accurately predicted.

Table 2. Example predictions with MIML-RBF.

No.	Ground truth	Predicted labels
1	UMA	UMA
2	LNA, LRI, UMA	LNA, LRA, UMA
3	LNA, UMA	LNA, UMA
4	LNA, LFX, LRA	LNA, LFX, LRI, LRA
5	LNA, LFX, LRA	LNA, LRA

5 Conclusion

In this study, we propose a novel method for the classification of multiple simultaneous frog species in environmental recordings. To the best of our knowledge, this is the first study that applies the MIML algorithm to frog calls. Since frogs tend to call simultaneously, the MIML algorithm is more suitable for dealing with those recordings than single-instance single-label classification. After applying acoustic event detection algorithm to each 10s recording, each frog syllable is segmented. Then, three feature sets are calculated based on those segmented

syllables. Finally, three MIML classifiers are used for the classification of frog calls with the best accuracy (81.8% true positive/negatives). Future work will focus on the study of novel features and MIML classifiers for further improving the classification performance.

References

1. J. Wimmer, M. Towsey, B. Planitz, I. Williamson, and P. Roe, "Analysing environmental acoustic data through collaboration and automation," *Future Generation Computer Systems*, vol. 29, no. 2, pp. 560–568, February 2013.
2. N. C. Han, S. V. Muniandy, and J. Dayou, "Acoustic classification of australian anurans based on hybrid spectral-entropy approach," *Applied Acoustics*, vol. 72, no. 9, pp. 639–645, 2011.
3. B. Gingras and W. T. Fitch, "A three-parameter model for classifying anurans into four genera based on advertisement calls," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 547–559, 2013.
4. C. Bedoya, C. Isaza, J. M. Daza, and J. D. López, "Automatic recognition of anuran species based on syllable identification," *Ecological Informatics*, vol. 24, pp. 200–209, 2014.
5. J. Xie, M. Towsey, A. Truskinger, P. Eichinski, J. Zhang, and P. Roe, "Acoustic classification of australian anurans using syllable features," in *2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (IEEE ISSNIP 2015)*, Singapore, Singapore, Apr. 2015.
6. Z.-H. Z. M.-L. Zhou, "Multi-instance multi-label learning with application to scene classification," *Advances in Neural Information Processing Systems*, p. 1609:1616, 2007.
7. F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4640–4650, 2012.
8. M.-L. Zhang and Z.-J. Wang, "Mimlrbf: Rbf neural networks for multi-instance multi-label learning," *Neurocomputing*, vol. 72, no. 16, pp. 3951–3956, 2009.
9. M.-L. Zhang, "A k-nearest neighbor based multi-instance multi-label learning algorithm," in *Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on*, vol. 2. IEEE, 2010, pp. 207–212.
10. P. Somervuo *et al.*, "Classification of the harmonic structure in bird vocalization," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 5. IEEE, 2004, pp. V–701.
11. C.-J. Huang, Y.-J. Yang, D.-X. Yang, and Y.-J. Chen, "Frog classification using machine learning techniques," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3737–3743, 2009.
12. M. Towsey, B. Planitz, A. Nantes, J. Wimmer, and P. Roe, "A toolbox for animal call recognition," *Bioacoustics*, vol. 21, no. 2, pp. 107–125, 2012.
13. J. Xie, M. Towsey, J. Zhang, and P. Roe, "Image processing and classification procedure for the analysis of australian frog vocalisations," in *Proceedings of the 2Nd International Workshop on Environmental Multimedia Retrieval*, ser. EMR '15. New York, NY, USA: ACM, 2015, pp. 15–20.
14. A. Mallawaarachchi, S. Ong, M. Chitre, and E. Taylor, "Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles,"

The Journal of the Acoustical Society of America, vol. 124, no. 2, pp. 1159–1170, 2008.

15. Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Miml: a framework for learning with ambiguous objects," *CORR abs/0808.3231*, 2008.
16. A. Dimou, G. Tsoumakas, V. Mezaris, I. Kompatsiaris, and I. Vlahavas, "An empirical study of multi-label learning methods for video annotation," in *Content-Based Multimedia Indexing, 2009. CBMI'09. Seventh International Workshop on*. IEEE, 2009, pp. 19–24.

Chapter 7

Detecting frog calling activity based on acoustic event detection and multi-label learning

---

7.1 Introduction

The publication in this chapter describes the research conducted for detecting frog calling activity (frog abundance and frog species richness). Different from chapter 6, acoustic event detection is used to predict the present or absence of eight specific frog species. In the mean time frog abundance is calculated based on the area and content of each segmented event. In chapter 6, acoustic features are calculated based on the results of AED to predict frog species richness, but the accuracy of AED results directly affect the MIML classification performance. To reduce the bias introduced by AED, this research presents a global feature representation for the classification of recordings with simultaneous vocalising frog species. This feature representation regards all the frog species in each individual recording as a whole. Therefore, the classification process can be framed as multiple-label (ML) learning.

Different from chapter 6, three global feature representations are extracted to classify each segmented recording: linear predictive coding, MFCCs and wavelet-based features. The wavelet-based features are similar with the features used in chapter 5. The only difference is that we divide *adaptive WPD sub-band ceptral coefficients* into three equal stages to capture more temporal information.

Further more, this proposed classification framework is conducted for a long-term analysis. The frog calling activity during the breeding season is calculated. Also, the correlation between

the frog calling activity and weather variables is studied.

7.2 Conference paper - Detecting frog calling activity based on acoustic event detection and multi-label learning

Procedia Computer Science

Volume 80, 2016, Pages 1–12

ICCS 2016. The International Conference on Computational Science

Detecting frog calling activity based on acoustic event detection and multi-label learning

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

Queensland University of Technology, Brisbane, Australia
j3.xie@hdr.qut.edu.au
{m.towsey, jinglan.zhang, p.roe}@qut.edu.au

Abstract

Frog population has been declining the past decade for habitat loss, invasive species, climate change, and so forth. Therefore, it is becoming ever more important to monitor the frog population. Recent advances in acoustic sensors make it possible to collect frog vocalizations over large spatio-temporal scale. Through the detection of frog calling activity with collected acoustic data, frog population can be predicted. In this paper we propose a novel method for detecting frog calling activity using acoustic event detection and multi-label learning. Here, frog calling activity consists of frog abundance and frog species richness, which denotes number of individual frog calls and number of frog species respectively. To be specific, each segmented recording is first transformed to a spectrogram. Then, acoustic event detection is used to calculate frog abundance. Meanwhile, those recordings without frog calls are filtered out. For frog species richness, three acoustic features, linear predictive coefficients, Mel-frequency Cepstral coefficients and wavelet-based features are calculated. Then, multi-label learning is used to predict frog species richness. Lastly, statistical analysis is used to reflect the relationship between frog calling activity (frog abundance and frog species richness) and weather variables. Experiment results show that our proposed method can accurately detect frog calling activity and reflect its relationship with weather variables.

Keywords: Frog abundance, frog species richness, multi-label learning, acoustic event detection, multiple regression analysis

1 Introduction

Over the past decade, a dramatic decline of frog population has been noticed worldwide [9]. Reasons for this decline can be summarized as habitat loss, invasive species, climate change. On one hand, frog population is rapidly declining, and on the other frogs are greatly important for the environment. First, frogs are an integral part of the food web, and the decline of their population can result in negative impacts through a whole-ecosystem. Second, frogs are

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

important indicator species for environmental health. Third, frogs are very useful in medical research that benefit humans[1]. Therefore, it is becoming necessary to protect frogs.

To monitor the change of frog population and optimize the protection policy, a growing number of researchers have shown interest in studying frogs [7, 3, 16]. Compared with counting frogs by visual observation, hearing frog vocalizations is much easier. Consequently, frogs' vocalizations are often used to study. Currently, there are two approaches for acosutic data collection. The traditional field survey method, which requires ecologists to physically visit sites for collecting bioacoustic data, is both time-consuming and costly [13]. Comparatively, recent advances in acoustic sensor techniques have greatly extended the spatio-temporal scale for collecting frog vocalizations. By the aid of acoustic sensors, it is possible to record frog vocalizations continuously and store them permanently. However, this technique provides us several gigabytes of compressed data per acoustic sensor every day. Develop automatic methods for exploring these quantities of acoustic data is becoming necessary.

Unfortunately, most previous methods focus on the recognition and classification of short-term recordings (segmented frog syllables) [7, 3, 16]. Few studies have explored long-term acoustic monitoring of frogs. Canavero *et al* [4] studied the relationship between calling activity of anuran assemblage and seasonal changes of weather factors such as temperature and rainfall. However, the activity of anuran species was simply quantified as rank abundance estimations of calling mate, which cannot accurately reflect frog species richness of each segmented recording. Ospina *et al* [10, 1] introduced a method named Region of Interest to identify the presence/absence of each species. Then, the detection model was built based on Hidden Markov Model with five frog parameters: maximum and minimum frequency, bandwidth, duration and slice between notes. The disadvantage of this method is that it can detecte the presence/absence of each frog species, but not frog abundance[2]. Akementins *et al* [2] explored the influence of abiotic cues on calling activity. Like [4], frog calling activity was quantified according to the numerical classification scheme. However, this method cannot accurately recognize frog species of each segmented recording. Xie *et al* [14] introduced acoustic indices for frog calling activity detection. Four indices, spectral peak track index, harmonic structure index, oscillation structure index, and Shannon entropy index, were selectively combined for detecting frog calling activity with a Gaussian mixture model. Then, the correlation between the frog calling activity and climate information was studied. However, the correlation was not strong because of the limitation of recording duration.

In this paper, we proposed a novel method for detecting frog calling activity. Here, frog calling activity, which consists of frog abundance and frog species richness, is detected based on acoustic event detection and multi-label learning. Frog abundance and frog species richness denote the number of individual frog calls and the number of different frog species of each segmented recording, respectively. Specifically, we first sample 10 seconds from every 10-minute recordings. Then, short-time Fourier transform (STFT) is used to obtain a spectrogram for each 10-second recording. Next, acoustic event detection is applied to the spectrogram image for frog abundance detection, which is also used to recognize those recordings without frog calls. Finally, multi-label learning is used to calculate frog species richness with three acoustic features: linear predictive coefficients, Mel-frequency Cepstral coefficients and wavelet-based features. After detecting frog abundance and frog species richness, statistical analysis is used to find the relationship between frog calling activity (frog abundance and frog species richness) and weather variables (temperature and rainfall). Experiment results show that our proposed method can accurately monitor frog calling activity and reflect its relationship with weather

[1] http://www.savethefrogs.com/why-frogs/
[2] Frog abundance denotes the number of individual frog calls

variables.

2 Materials and methods

The architecture of our calling activity detection system is shown in Figure 1. The system consists of three parts: frog abundance detection, frog species richness detection, and correlation analysis.

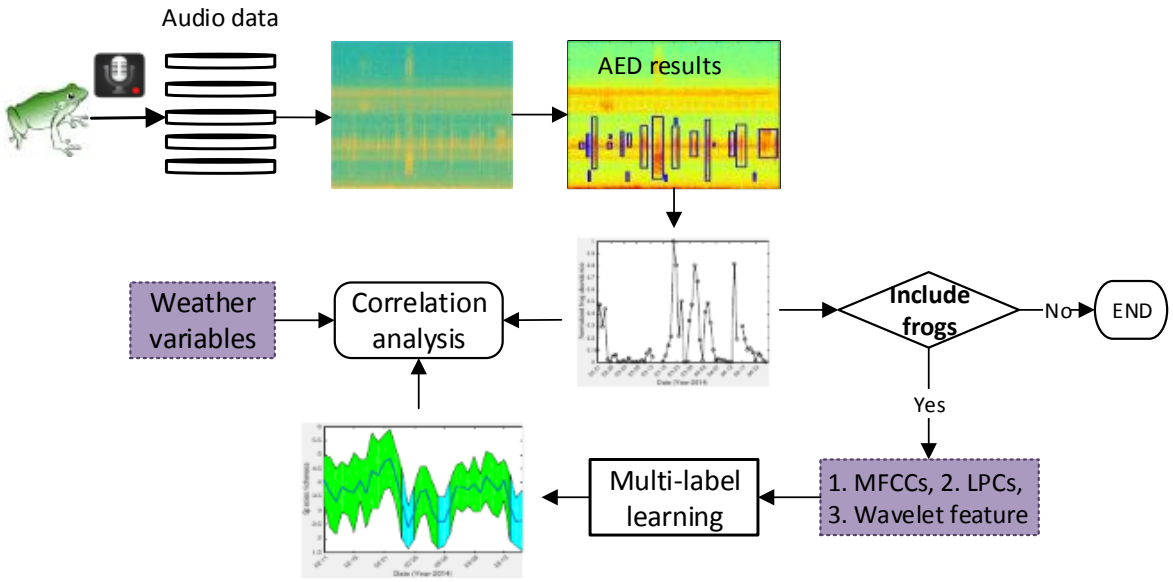


Figure 1: Flowchart of frog call classification system

2.1 Acquisition of frog call recordings

All recordings selected for this study were obtained from three sites in Queensland, Australia: *Kiyomi dam*, *Stony creek dam* and *BG creek dam*, using a battery-powered acoustic sensor (stored in a weather proof metal box) with an external microphone. The recordings were stored on 16GB SD cards in 64 kbps MP3 mono format. All recordings were collected from February, 2014 to April, 2014, because it is the breeding season in Queensland when male frogs make calls to attract female for reproducing. All recordings started around sunset, finished around sunrise every day and have 12 hour duration. We sampled 10-second recordings every 10 minutes for those continuous recordings. There are 4170, 4908, and 1544 10-second recordings for *Kiyomi dam*, *Stony creek dam* and *BG creek dam* respectively, because of data loss. A representative sample of 342 10-second recordings was selected to train and evaluate the proposed method. The ground truth of those 342 10-second recordings is generated by a frog expert who manually tags each recording with frog species.

We first manually inspected spectrograms of ten randomly selected call examples for each frog species. Two parameters, dominant frequency and syllable duration, were then measured and averaged, as listed in Table 1, which are used as prior information for subsequent analysis.

Table 1: Dominant frequency (F_0) and syllable duration (T_s) of eight frog species averaged for ten randomly selected syllables.

Frog species	Code	Dominant frequency (Hz)	Syllable duration(ms)
Canetoad	CAD	560	NA
Cyclorana novaehollandiae	CNE	610	400
Limnodynastes terraereginae	LTE	610	100
Litoria fallax	LFX	4000	280
Litoria nasuta	LNA	2800	160
Litoria rothii	LRI	1800	500
Litoria rubella	LRA	2300	580
Uperolela mimula	UMA	2400	120

2.2 Frog abundance monitoring

Frog abundance is monitored through the detection of acoustic events in a spectrogram image. Here, the spectrogram was generated by applying short-time Fourier transform (STFT) to each 10-second recording. Acoustic event detection, which consists of multiple image processing steps, are modified from our previous study [15] and summarized as follows.

Step 1: Wiener filter

To de-noise and smooth the spectrogram, a 2-dimensional Wiener filter is applied to the spectrogram image over a 5×5 time-frequency grid, where the filter size is selected after the consideration of trade-off between removing the background graininess and blurring acoustic events.

$$\hat{S_{tf}} = \mu + \frac{(\sigma^2 - \nu^2)}{\sigma^2}(S_{tf} - \nu) \tag{1}$$

where μ and σ^2 are local mean and variance, respectively. ν^2 is the noise variance estimated by averaging all local variances.

Step 2: Spectral subtraction

After Wiener filtering, the graininess has been removed. However, some noises such as wind, insect, motor engine that cover the whole recording are still remained. Here, a modified spectral subtraction is employed for dealing with those noises. Description of this algorithm can be found in our previous study [16].

Step 3: Adaptive thresholding

Following noise reduction, the next step is to convert the noise reduced spectrogram $\hat{S'_{tf}}$ into the binary spectrogram S^b_{tf} for events detection. Here, an adaptive thresholding method named *Otsu thresholding* [11] is employed to find an optimal threshold.

$$\phi_b^2(k) = w_1(k)w_2(k)[\mu_1(k) - \mu_2(k)]^2 \tag{2}$$

where $w_1(k) = \sum_0^k p(j)$ is calculated from the histogram as k, $p(j) = n(j)/N$ are the values of the normalized gray level histogram, $n(j)$ is the number of values in level j, N is the total number of values over the whole spectrogram image, $\mu_1(k) = [\sum_0^k p(j)x(j)]/w_1$, $x(j)$ is the value at the center of the jth histogram bin. Then, the threshold, T_0, is calculated as

$$T_0 = (\phi_{b1}^2(k) + \phi_{b2}^2(k))/2 \tag{3}$$

Step 4: Events filtering using dominant frequency and event area

To further remove those events that are not belong to frog species shown in Table 1, dominant frequency (F_0) and area (number of pixels) within the event boundary (Ar) are used for filtering. First, large acoustic events, whose area is larger than A_{large}, are separated into small events, because the area of frog calls to be classified in Table 1 is empirically smaller than A_{large}. Then,

dominant frequency is used to filter the events. First, the averaged frequency is calculated by averaging the peak frequency within each acoustic event. Then, the event, whose averaged frequency are not within allowed fluctuation in both sides of dominant frequency, are discarded. Finally, small acoustic events, whose area is smaller than A_{small}, are filtered. Those events, whose average frequency are between 300 Hz and 800 Hz, are not filtered using A_{small}, because the area of LTE (averaged frequency is between 300 Hz and 800 Hz) is smaller than A_{small}. Figure 2 shows the acoustic event detection results.

To calculate frog abundance of each segmented recording, the frog abundance is calculated as follows.

$$F_{abun} = \sum_{n=1}^{N} A_{i,j}(n)^2 \tag{4}$$

Here, $A_{i,j}$ represents the decibel value of location (i, j) within each acoustic event n in the spectrogram.

2.3 Wavelet-based feature extraction for species richness analysis

Frog species richness is calculated by tagging each segmented recording. Since many segmented recordings consist of multiple frog species, one direct solution is to assign each recording with a set of labels (frog species) for explicitly expressing its semantics [17]. Therefore, multi-label learning is adopt to tag each segmented recording.

Extracting discriminating features, which maximize between-group (inter-specie) dissimilarity and minimize within-group (intra-specie) dissimilarity, is very important for achieving high classification performance [7, 3]. In this study, feature extraction is performed based on wavelet packet decomposition using a modified version of the method introduced in [16] and summarized below.

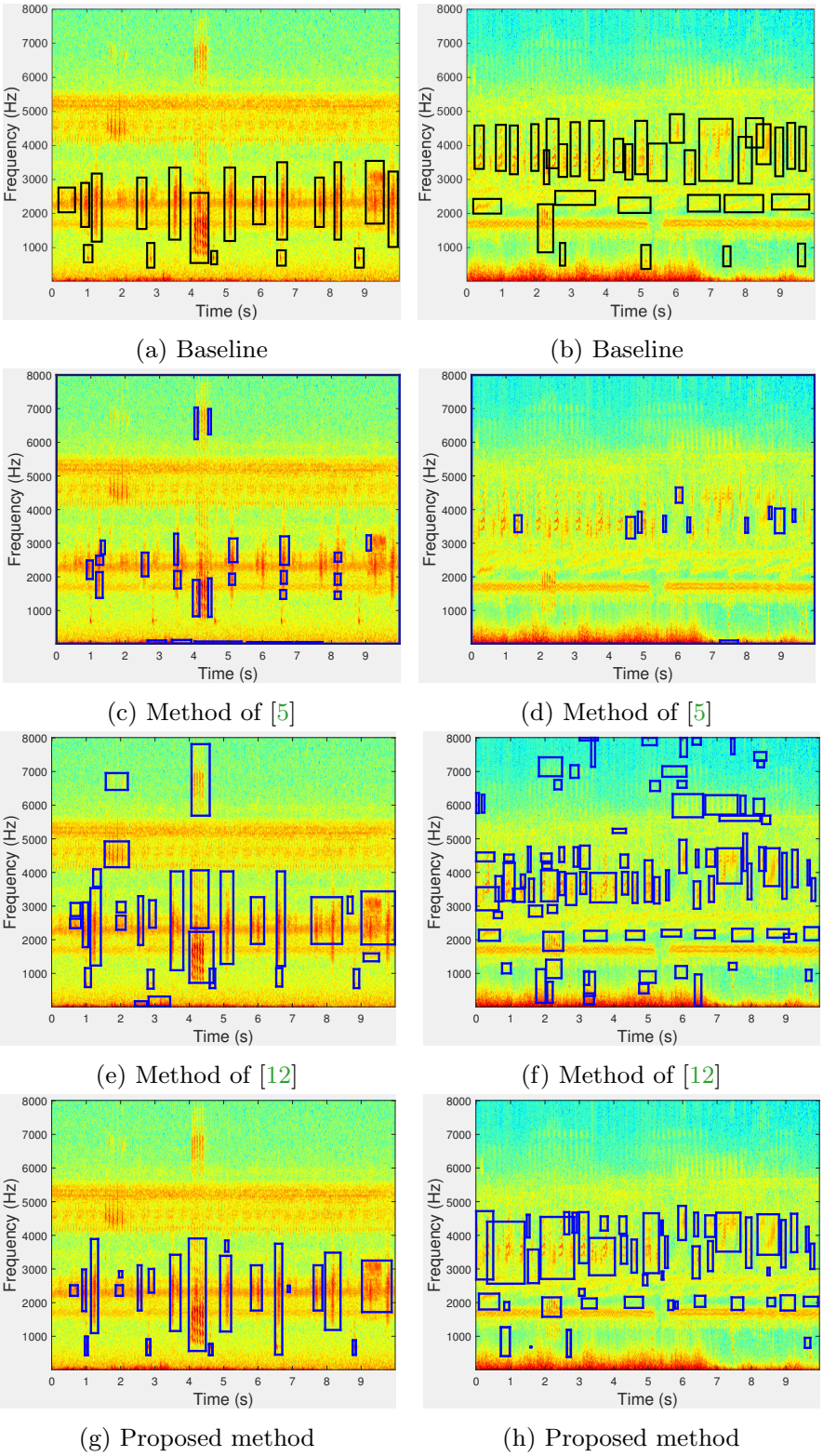
For feature extraction, constructing a suitable frequency scale for wavelet packet (WP) tree based on the dominant frequency of each frog species is the first step, because different frog species tend to have different dominant frequencies [6]. In [16], k-means clustering was first applied to the extracted dominant frequencies of training data. Then, the frequency scale was built by sorting clustering centroids to construct the WP tree. In this study, the prior information (F_0) obtained from Table 1 is directly used to construct the WP tree. We iteratively detect each WP tree sub-band node until the frequency range of each node includes more than one F_0. Then, the WP tree of that particular sub-band node will be further split until each sub-band node has only one dominant frequency value or none. After constructing the frequency scale, adaptive frequency scaled wavelet packet decomposition is applied to each segmented recording for feature extraction.

For each 10-second recording, it is represented as $y(n)$, $n = 1, ..., N$, where N is the length of each recording. Based on the $y(n)$, detailed description for WP-based feature extraction is list as follows:

Step 1: Add a Hamming window to the signal y(n) and perform wavelet packet decomposition spaced in adaptive frequency scale as described in [16].

$$WP(i, j) = \sum_{i=1}^{M} y(n)w(n)\psi_{(a,b)}(n) \tag{5}$$

where $w(L)$ is the Hamming window function, $WP(i, j)$ is the wavelet coefficients of the decomposition, i is the sub-band index, j is the index of wavelet coefficients, $\psi_{(a,b)}(n)$ is the



(a) Baseline (b) Baseline

(c) Method of [5] (d) Method of [5]

(e) Method of [12] (f) Method of [12]

(g) Proposed method (h) Proposed method

Figure 2: Acoustic event detection for frog abundance monitoring using different methods. For each row, different methods are applied to the same recordings. The baseline of the detection results is show in the first row; detected frog calls are draw using a blue rectangle.

wavelet base function, and we use 'db4' experimentally. Here, a and b are the scale and shift parameters, respectively.

Step 2: Calculate the total energy of each sub-band.

$$WP_i = \sum_{j=1}^{M_i} [WP(i,j)]^2 \tag{6}$$

where $i = 1, 2, ..., T$, and T is the total number of sub-band, and $j = 1, 2, ..., M_i$, M_i is the total number of wavelet coefficients.

Step 3: Normalize the energy of each sub-band.

$$SE_i = \frac{WP_i}{M_i} \tag{7}$$

where $i = 1, 2, ..., T$.

Step 4: Perform discrete cosine transform on the logarithm sub-band energy for dimension reduction and obtain the WP-based feature.

$$WP_{base}(d) = \sum_{i=1}^{T} logSE_i cos(\frac{d(i-0.5)}{T}\pi) \tag{8}$$

where $d = 1, 2, ..., d^{'}$, $1 \leq d^{'} \leq T$, here $d^{'}$ is the dimension of WP-based feature, and set as 12.

Different from [16], the recording is first segmented into frames using a Hamming window. Then, all frames are divided into three equal parts, and WP-feature within each part is averaged, respectively, because different frog species within similar frequency band may exist in one 10-second recording, segmenting each recording into small parts might be able to keep the information of different frog species in the same frequency band. Besides WP-based feature, two other acoustic features, linear predictive coefficients (LPCs) and Mel-frequency Cepstral coefficients (MFCCs), are also calculated for the comparison.

2.4 Multi-label classification for species richness analysis

Since many segmented recordings consist of calls from multiple frog species, frog call classification can be framed as a multi-label classification problem. However, previous studies have not adopted multi-label learning to classify frog calls. Therefore, it is worth to investigate different multi-label learning algorithms for the classification of multiple vocalizing frog species. In this study, four multi-label learning algorithms, whose base classifier is C4.5 decision tree, are employed: binary relevance (BR), classifier chains (CC), random k-labEL Pruned Sets (RAKEL and RAKEL1) [17]. The default parameter settings of those four multi-label learning algorithms are used. The trained classifier, which achieves the best classification performance, is then used to tag rest recordings. After tagging each 10-second recording, frog species richness is lastly calculated as follows.

$$F_{rich} = \frac{\sum_{k=1}^{K} f_{rich}(k)}{K} \tag{9}$$

where $f_{rich}(k)$ is the number of frog species of each tagged 10-second recording, K is the number of 10-second recording for each day.

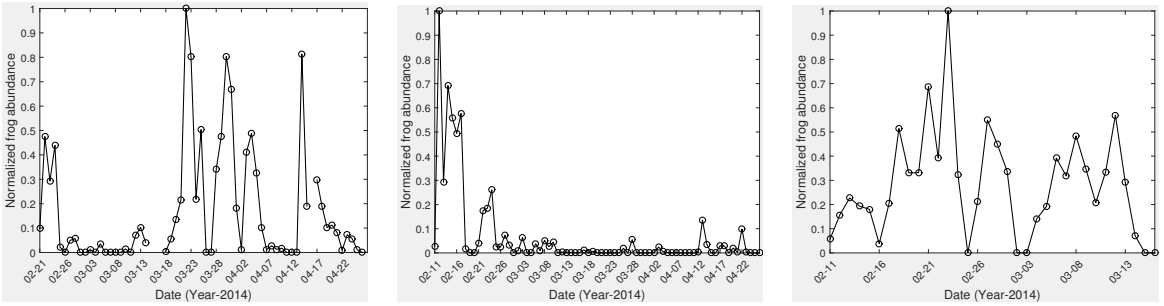
3 Experiment setup

Each 10-second recording is divided into frames of 512 samples and 50% frame overlap for STFT. A_{large} and A_{small}, which are used for area filtering in acoustic event detection, are empirically set at 3000 pixels and 300 pixels, respectively. Allowed fluctuations in both sides of dominant frequency are 300 Hz for dominant frequency filtering. For WP-based feature, window size and overlap are 512 samples and 50%, the window function is a Hamming window. All algorithms were programmed in Matlab 2014b except multi-label learning, which was implemented in Meka 1.7.7[4].

4 Experiment results

4.1 Frog abundance detection

Figure 3 shows the frog abundance result of three selected sites through the whole frog breeding season. It can be found that the frog abundance of the same site changes a great deal over time. In the *Kiyomi dam*, frog abundance is relatively high from February 21 to February 25. However, frog abundance is quite low in two period, which are February 26 to March 11 and April 07 to April 12. The highest abundance of this site is achieved on March 22. However, the highest abundance for *Stony creek dam* and *BG creek dam* is obtained in February, which shows that frog abundance of different sites often varies a lot for different environments. Recordings of 47 days of all three sites have no frog calls. In the subsequent analysis, only those recordings that consist of frog calls are used for frog species richness analysis.



(a) *Kiyomi dam* (b) *Stony creek dam* (c) *BG creek dam*

Figure 3: Frog abundance detection of different sites: *Kiyomi dam*, *Stony creek dam* and *BG creek dam*. For *Kiyomi dam*, three days do not record any acoustic data and then there is no value in those particular days. All the frog abundance value is normalized to [0 1].

4.2 Frog species richness analysis

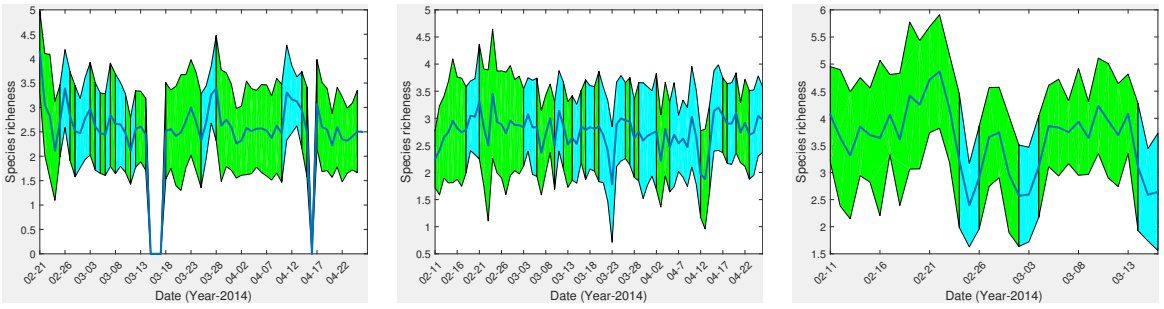
We apply different multi-label learning algorithms on 342 selected recordings to compare different feature sets. Then, six evaluation rules are used to compare the performance with the combination of four feature sets and four multi-label algorithms: Hamming loss, Rank loss, Average precision, One error, Exampled based F1, and Micro F1 [8, 17]. Experiment results are shown in Table 2.

[4]http://meka.sourceforge.net/

Table 2: Classification results based on four feature sets and four multi-label learning algorithms. Here the methods for multi-label algorithms are in accordance to the name in the Meka software. The base classifier of all methods is decision tree. For a metric, the best value is in bold. Here, ↓ indicates the smaller the better, while ↑ indicates the bigger the better.

Features	Method	Hamming loss ↓	Rank loss ↓	Average precision ↑	One error ↓	Exampled based F1 ↑	Micro F1 ↑
MFCCs+LPCs	BR	0.155 ± 0.015	0.171 ± 0.037	**0.446 ± 0.061**	0.246 ± 0.063	0.699 ± 0.03	0.749 ± 0.024
	CC	0.147 ± 0.018	0.147 ± 0.02	0.35 ± 0.016	0.199 ± 0.042	0.722 ± 0.035	0.756 ± 0.029
	RAkEL	0.167 ± 0.038	0.122 ± 0.026	0.333 ± 0.017	0.194 ± 0.063	0.721 ± 0.044	0.752 ± 0.041
	RAkEL1	0.134 ± 0.012	0.099 ± 0.025	0.342 ± 0.023	**0.147 ± 0.056**	0.74 ± 0.044	0.783 ± 0.022
Multi-stage MFCCs + LPCs	BR	0.155 ± 0.016	0.169 ± 0.035	0.445 ± 0.062	0.249 ± 0.064	0.7 ± 0.03	0.75 ± 0.024
	CC	0.147 ± 0.018	0.147 ± 0.021	0.35 ± 0.016	0.199 ± 0.042	0.722 ± 0.034	0.756 ± 0.028
	RAkEL	0.166 ± 0.035	0.124 ± 0.027	0.334 ± 0.018	0.194 ± 0.069	0.724 ± 0.048	0.754 ± 0.04
	RAkEL1	0.134 ± 0.013	0.101 ± 0.026	0.342 ± 0.02	0.15 ± 0.063	0.737 ± 0.05	0.783 ± 0.023
WP-based feature + LPCs	BR	0.148 ± 0.025	0.139 ± 0.033	0.356 ± 0.065	0.254 ± 0.063	0.708 ± 0.046	0.762 ± 0.036
	CC	0.168 ± 0.031	0.168 ± 0.045	0.341 ± 0.027	0.272 ± 0.061	0.684 ± 0.054	0.723 ± 0.048
	RAkEL	0.155 ± 0.023	0.103 ± 0.022	0.324 ± 0.018	0.178 ± 0.031	0.729 ± 0.032	0.763 ± 0.030
	RAkEL1	0.14 ± 0.027	0.094 ± 0.018	0.333 ± 0.028	0.193 ± 0.063	0.727 ± 0.053	0.773 ± 0.042
Multi-stage WP-based feature + LPCs	BR	0.153 ± 0.014	0.147 ± 0.022	0.364 ± 0.056	0.266 ± 0.037	0.689 ± 0.035	0.75 ± 0.025
	CC	0.142 ± 0.029	0.146 ± 0.023	0.345 ± 0.019	0.254 ± 0.094	0.714 ± 0042	0.764 ± 0.045
	RAkEL	0.154 ± 0.022	0.11 ± 0.012	0.33 ± 0.027	0.196 ± 0.062	0.739 ± 0.022	0.768 ± 0.025
	RAkEL1	**0.131 ± 0.012**	**0.09 ± 0.014**	0.33 ± 0.026	0.173 ± 0.03	**0.743 ± 0.026**	**0.787 ± 0.018**

The combination of multi-stage WP-based feature+LCPs and the RAkEL1 method achieves the best performance. Therefore, this combination is used for the testing data. Figure 4 shows the frog species richness of the three selected sites. For all the three sites, the variation of species richness is not high, which shows that species richness of the same area is relatively stable. However, frog species richness of *BG creek dam* has a smaller variation over the time than *Kiyomi dam* and *Stony creek dam*. The comparison of the species richness for the three sites is shown in Figure 5. In contrast to other sites, the species richness in *BG creek dam* is the highest. This might be that *BG creek dam* is closer to a river and farther away from the human community.



(a) *Kiyomi dam* (b) *Stony creek dam* (c) *BG creek dam*

Figure 4: Frog species richness distribution of three selected sites. Here green bar represents the species variation, blue bar means there is no frog calls, zero value denotes the data loss of those particular days.

4.3 Statistical analysis

Multiple regression analysis is used to explore frog calling activity (frog abundance and frog species richness) along weather variables (mean temperature and rainfall)[5]. Frog calling activity is found to be highly correlated with mean temperature ($F=5.18$, $P<0.05$ for abundance, and $F=10.7$, $P<0.01$ for species richness). To calculate the correlation between rainfall and

[5]http://www.bom.gov.au/?ref=hdr

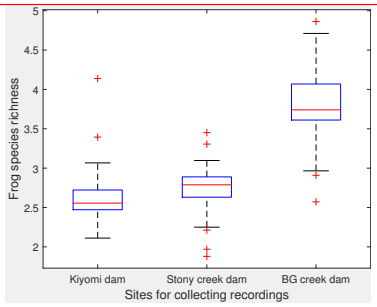


Figure 5: Averaged frog species richness of different sites.

frog calling activity, we first set the rainfall vaule as the dummy variable. Then, the correlation between frog calling activity and rainfall is also studied with multiple regression analysis ($F=4.63$, $P<0.05$ for abundance, and $F=4.64$, $P<0.05$ for species richness). The statistical analysis results indicate that frogs tend to make calls in the warm and humidity environment, which is in accordance to previous studies [2, 4].

5 Conclusion and future work

Acoustic sensors are more widely used to monitor frog calling activity than the traditional field survey method. However, the use of acoustic sensors generates large volumes of audio data, which makes it necessary to develop automated methods. This paper proposes a novel method for detecting frog calling activity based on acoustic event detection and multi-label learning. Specifically, acoustic event detection is the first step to calculate frog abundance. Meanwhile, each 10-second recording is analyzed to decide whether it has frog calls or not. For those recordings with frog calls, multi-label learning is further used for calculating frog species richness with multi-stage WP-based features and LPCs. Finally, statistic analysis is utilized to reflect the relationship between frog calling activity (frog abundance and frog species richness) and weather variables (mean temperature and rainfall). Experiment results show that our proposed method can accurately detect frog calling activity and reflect its relationship with weather variables. Future work will focus on a wider frog call database, including a larger number of frog species, and frog calls collected over a longer period.

6 Acknowledgement

The authors would like to thank Mingying Zhu for the help of statistic analysis, the support of James Cook University, China Scholarship Council and Wet Tropics Management Authority.

References

[1] T Mitchell Aide, Carlos Corrada-Bravo, Marconi Campos-Cerqueira, Carlos Milan, Giovany Vega, and Rafael Alvarez. Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1:e103, 2013.

[2] MS Akmentins, LC Pereyra, EA Sanabria, and M Vaira. Patterns of daily and seasonal

calling activity of a direct-developing frog of the subtropical andean forests of argentina. *Bioacoustics*, 24(2):89–99, 2015.

[3] Carol Bedoya, Claudia Isaza, Juan M Daza, and José D López. Automatic recognition of anuran species based on syllable identification. *Ecological Informatics*, 24:200–209, 2014.

[4] Andrés Canavero, Matías Arim, Daniel E Naya, Arley Camargo, Inés Da Rosa, and Raúl Maneyro. Calling activity patterns in an anuran assemblage: the role of seasonal trends and weather determinants. *North-Western Journal of Zoology*, 4(1):29–41, 2008.

[5] Gábor Fodor. The ninth annual mlsp competition: first place. In *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*, pages 1–2. IEEE, 2013.

[6] Bruno Gingras and William Tecumseh Fitch. A three-parameter model for classifying anurans into four genera based on advertisement calls. *The Journal of the Acoustical Society of America*, 133(1):547–559, 2013.

[7] Chenn-Jung Huang, Yi-Ju Yang, Dian-Xiu Yang, and You-Jia Chen. Frog classification using machine learning techniques. *Expert Systems with Applications*, 36(2):3737–3743, 2009.

[8] Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sao Deroski. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084 – 3104, 2012.

[9] Malcolm L McCallum. Amphibian decline or extinction? current declines dwarf background extinction rate. *Journal of Herpetology*, 41(3):483–491, 2007.

[10] Oscar E Ospina, Luis J Villanueva-Rivera, Carlos J Corrada-Bravo, and T Mitchell Aide. Variable response of anuran calling activity to daily precipitation and temperature: implications for climate change. *Ecosphere*, 4(4):art47, 2013.

[11] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[12] Michael W. Towsey and Birgit Planitz. Technical report : acoustic analysis of the natural environment. 2011, April 2011.

[13] Jie Xie, Michael Towsey, Anthony Truskinger, Philip Eichinski, Jinglan Zhang, and Paul Roe. Acoustic classification of australian anurans using syllable features. In *2015 IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (IEEE ISSNIP 2015)*, Singapore, Singapore, April 2015.

[14] Jie Xie, Michael Towsey, Kiyomi Yasumiba, Jinglan Zhang, and Paul Roe. 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, April 2015.

[15] Jie Xie, Michael Towsey, Jinglan Zhang, and Paul Roe. Image processing and classification procedure for the analysis of australian frog vocalisations. In *Proceedings of the 2Nd International Workshop on Environmental Multimedia Retrieval*, EMR '15, pages 15–20, New York, NY, USA, 2015. ACM.

[16] Jie Xie, Michael Towsey, Jinglan Zhang, and Paul Roe. Adaptive frequency scaled wavelet packet decomposition for frog call classification. *Ecological Informatics*, 32:134 – 144, 2016.

[17] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, 26(8):1819–1837, Aug 2014.

Chapter 8

Conclusions and Future work

In this thesis, both feature extraction methods and classification frameworks for frog call classification are investigated. For feature extraction, enhanced feature representations and a novel feature representation based on wavelet packet decomposition are proposed. As for the classification framework, we adopt both MIML learning and ML learning frameworks to study multiple simultaneously vocalising frog species in low SNR recordings.

Many challenges of this thesis lies in the designing and identifying the effective feature extraction algorithms and adopting novel classification frameworks that can successfully classify low SNR recordings with multiple simultaneously vocalising frog species. In this chapter, key contributions of this research to the challenge will be summarised. Furthermore, useful avenues of inquiry for improving the methods described in this thesis will be explored.

8.1 Summary of contributions

Four contributions have been made by this research:

- *Enhanced acoustic feature representation for frog call classification in high SNR recordings.*

  The classification of frog calls has been addressed in this thesis using both high and low SNR recordings. A systematic scheme was developed towards the goal of automatic classification of frog calls. The performances of various classification methods such as LDA, K-NN, SVM, RF, MLP were evaluated together with different feature representations. The experience gained and experimental results demonstrate that: 1) Compared with

previous feature representation, an enhanced feature representation including temporal, perpetual, and cepstral features can achieve the best classification performance. 2) The best classification performance is achieved by SVM and RF, in comparison with LDA, K-NN, and MLP. 3) The cepstral features are very sensitive to the background noise, but can achieve good classification accuracy in the high SNR recordings.

- *A novel feature representation via WPD for frog call classification in both high and low SNR recordings.*

  To improve the anti-noise ability of cepstral features, wavelet packet decomposition is utilised to design a novel cepstral feature representation. Compared with other cepstral features such as MFCCs, Mel-scale wavelet packet decomposition coefficients, our proposed feature representation shows both good classification performance and excellent anti-noise ability.

- *Design a MIML classification framework for frog call classification in low SNR recordings.*

  Since most frog field recordings consist of multiple simultaneously vocalising frog species, both MIML and ML classification frameworks are first introduced to study frog calls. For MIML learning, a novel acoustic event detection algorithm is designed to segment acoustic events by using events filtering. Then, different MIML classifiers are evaluated together with various acoustic features based on the content and shape of the segmented events. The results show that MIML-RBF achieves the best classification results.

- *Design a ML classification framework for frog call classification in low SNR recordings.*

  For a ML classification framework, acoustic event detection is first used to filter all the recordings to find those that have frog calls. Meanwhile, frog abundance is detected based on the shape and content of segmented acoustic events. Then, those recordings with frog calls are classified via ML learning. The feature representation used is a modified adaptive WPD sub-band cepstral coefficients. Compared with MIML learning, ML learning can achieve a better classification performance. Lastly, the correlation between the frog calling activity (frog abundance and frog species richness) and weather variables (mean temperature and rainfall) are studied.

8.2 Limitations and further research

On the topic of this thesis, there is still much work that can be done to help scientists and researchers in data collection and analysis in the bioacoustics communities. One of the most important issues when dealing with frog recordings is the need for the existence of standardised species-specific data with behavioural labels. Therefore, the algorithms we developed for frog call classification can be evaluated on a larger dataset. Consequently, researchers will be able to use the outcomes of such automatic call classification methods for field studies.

Another aspect which requires tremendous improvement is the need for an advanced frog syllable segmentation method for the field recordings so as to extract more accurate event-based features and conduct more thorough analysis on frog vocalisations. The problem of syllable segmentation is very complicated, because there are many simultaneous overlapping calling activities from birds, frogs, insects, and many other sources. In addition, the *adaptive WPD sub-band cepstral coefficients* feature has been successfully used for frog call classification, which is used to capture the frequency domain information. The time-varying information still has not been explicitly addressed for frog call classification.

Our developed frog call classification system aims to help ecologists to study frogs over larger spatial and temporal scales. However, there is still no a generic platform for running the frog calls recordings. It is necessary to develop an on-line website with our developed frog call classification algorithms, and then ecologists can do the analysis on their own. Another important aspect of practical systems is the speed of data processing executed through classification algorithms. For this purpose, the MATLAB code corresponding to feature extractors and classifiers needs to be optimised to perform real-time frog call classification in the field.

Appendix A

Frog species studied in this study

Literature Cited

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

Allen, J. (1977). Short term spectral analysis, synthesis, and modification by discrete fourier transform. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 25(3):235–238.

Allen, J. (1997). Short term spectral analysis, synthesis, and modification by discrete fourier transform. In *IEEE Trans. on Acoust., Speech, and Sig. Proc.*, volume 4, pages 21–24.

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

Boashash, B. and Black, P. (1987). An efficient real-time implementation of the wigner-ville distribution. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(11):1611–1618.

Brandes, T. S. (2008). Feature vector selection and use with hidden markov models to identify frequency-modulated bioacoustic signals amidst noise. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(6):1173–1180.

Brandes, T. S., Naskrecki, P., and Figueroa, H. K. (2006). Using image processing to detect and classify narrow-band cricket and frog calls. *The Journal of the Acoustical Society of America*, 120(5):2950–2957.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

Camacho, A., García-Rodríguez, A., and Bolaños, F. (2011). Automatic detection of vocalizations of the frog diasporus hylaeformis in audio recordings. In *Proceedings of Meetings on Acoustics*, volume 14, page 010003. Acoustical Society of America.

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

Colombia, C. and del Cauca, V. (2009). Frogs species classification using lpc and classification algorithms on wireless sensor network platform.

Colonna, J., Ribas, A., dos Santos, E., and Nakamura, E. (2012). Feature subset selection for automatically classifying anuran calls using sensor networks. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8.

Colonna, J. G., Cristo, M., Junior, M. S., and Nakamura, E. F. (2015). An incremental technique for real-time bioacoustic signal segmentation. *Expert Systems with Applications*, 42(21):7367 – 7374.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Croker, B. and Kottege, N. (2012). Using feature vectors to detect frog calls in wireless sensor networks. *The Journal of the Acoustical Society of America*, 131(5):EL400–EL405.

Dang, T., Bulusu, N., and Hu, W. (2008). Lightweight acoustic classification for cane-toad monitoring. In *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pages 1601–1605. IEEE.

Dayou, J., Han, N. C., Mun, H. C., Ahmad, A. H., Muniandy, S. V., and Dalimin, M. N. (2011). Classification and identification of frog sound based on entropy approach. In *International Conference on Life Science and Technology*, volume 3, pages 184–187.

Deng, L. and O'Shaughnessy, D. (2003). *Speech processing: a dynamic and optimization-oriented approach*. CRC Press.

Dorcas, M. E., Price, S. J., Walls, S. C., and Barichivich, W. J. (2009). Auditory monitoring of anuran populations. *Amphibian ecology and conservation: a hand book of techniques. Oxford University Press, Oxford*, pages 281–298.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z.-I., Knowler, D. J., Lévêque, C., Naiman, R. J., Prieur-Richard, A.-H., Soto, D., Stiassny, M. L., et al. (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological reviews*, 81(02):163–182.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Fox, E. J. (2008). A new perspective on acoustic individual recognition in animals with limited call sharing or changing repertoires. *Animal Behaviour*, 75(3):1187 – 1194.

Gingras, B. and Fitch, W. T. (2013). A three-parameter model for classifying anurans into four genera based on advertisement calls. *The Journal of the Acoustical Society of America*, 133(1):547–559.

Gordon, L., Chervonenkis, A. Y., Gammerman, A. J., Shahmuradov, I. A., and Solovyev, V. V. (2003). Sequence alignment kernel for recognition of promoter regions. *Bioinformatics*, 19(15):1964–1971.

Grigg, G., Taylor, A., Mc Callum, H., and Watson, G. (1996). Monitoring frog communities: an application of machine learning. In *Proceedings of Eighth Innovative Applications of Artificial Intelligence Conference, Portland Oregon*, pages 1564–1569.

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

Ho, T. K. (1995). Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification.

Huang, C.-J., Chen, Y.-J., Chen, H.-M., Jian, J.-J., Tseng, S.-C., Yang, Y.-J., and Hsu, P.-A. (2014). Intelligent feature extraction and classification of anuran vocalizations. *Applied Soft Computing*, 19(0):1 – 7.

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

Huang, C.-J., Yang, Y.-J., Yang, D.-X., Chen, Y.-J., and Wei, H.-Y. (2008). Realization of an intelligent frog call identification agent. In *Agent and Multi-Agent Systems: Technologies and Applications*, pages 93–102. Springer.

Itakura, F. (1975). Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, 57(S1):S35–S35.

Jaafar, H. and Ramli, D. (2013). Automatic syllables segmentation for frog identification system. In *Signal Processing and its Applications (CSPA), 2013 IEEE 9th International Colloquium on*, pages 224–228.

Jaafar, H., Ramli, D. A., and Shahrudin, S. (2013a). A comparative study of classification algorithms and feature extractions for frog identification system.

Jaafar, H., Ramli, D. A., and Shahrudin, S. (2013b). Mfcc based frog identification system in noisy environment. In *Signal and Image Processing Applications (ICSIPA), 2013 IEEE International Conference on*, pages 123–127. IEEE.

Jang, Y., Hahm, E. H., Lee, H.-J., Park, S., Won, Y.-J., and Choe, J. C. (2011). Geographic variation in advertisement calls in a tree frog species: gene flow and selection hypotheses. *PloS one*, 6(8):e23297.

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

Lek, S. and Guégan, J.-F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological modelling*, 120(2):65–73.

Meyer, Y. and Salinger, D. H. (1995). *Wavelets and operators*, volume 1. Cambridge university press.

Noda, J. J., Travieso, C. M., and S?nchez-Rodr?guez, D. (2016). Methodology for automatic bioacoustic classification of anurans based on feature fusion. *Expert Systems with Applications*, 50:100 – 106.

Potamitis, I. (2015). Unsupervised dictionary extraction of bird vocalisations and new tools on assessing and visualising bird activity. *Ecological Informatics*, 26:6–17.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Rocchesso, D. (2003). *Introduction to sound processing*. Mondo estremo.

Samarasinghe, S. (2006). *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition*. CRC Press.

Somervuo, P., Härmä, A., and Fagerlund, S. (2006). Parametric representations of bird sounds for automatic species recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(6):2252–2263.

Stewart, D. (1999). Australian frog calls: subtropical east. Audio CD.

Tan, W., Jaafar, H., Ramli, D., Rosdi, B., and Shahrudin, S. (2014). Intelligent frog species identification on android operating system. *International journal of circuits, systems and signal processing*.

Tjahja, T. V., Fern, X. Z., Raich, R., and Pham, A. T. (2015). Supervised hierarchical segmentation for bird song recording. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 763–767. IEEE.

Vaca-Castano, G. and Rodriguez, D. (2010). Using syllabic mel cepstrum features and k-nearest neighbors to identify anurans and birds species. In *Signal Processing Systems (SIPS), 2010 IEEE Workshop on*, pages 466–471. IEEE.

Wei, B., Yang, M., Rana, R. K., Chou, C. T., and Hu, W. (2012). Distributed sparse approximation for frog sound classification. In *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, pages 105–106. ACM.

Wimmer, J., Towsey, M., Planitz, B., Williamson, I., and Roe, P. (2013). Analysing environmental acoustic data through collaboration and automation. *Future Generation Computer Systems*, 29(2):560–568.

Xie, J., Towsey, M., Eichinski, P., Zhang, J., and Roe, P. (2015a). Acoustic feature extraction using perceptual wavelet packet decomposition for frog call classification. *11th IEEE International Conference on eScience*.

Xie, J., Towsey, M., Truskinger, A., Eichinski, P., Zhang, J., and Roe, P. (2015b). Acoustic classification of australian anurans using syllable features. In *2015 IEEE Tenth International*

Conference on Intelligent Sensors, Sensor Networks and Information Processing (IEEE ISSNIP 2015), Singapore, Singapore.

Xie, J., Towsey, M., Zhang, J., Dong, X., and Roe, P. (2015c). Application of image processing techniques for frog call classification. *International Conference on Image Processing*.

Xie, J., Towsey, M., Zhang, J., and Roe, P. (2015d). Image processing and classification procedure for the analysis of australian frog vocalisations. In *Proceedings of the 2Nd International Workshop on Environmental Multimedia Retrieval*, EMR '15, pages 15–20, Shanghai, China. ACM.

Xie, J., Towsey, M., Zhang, J., and Roe, P. (2016). Adaptive frequency scaled wavelet packet decomposition for frog call classification. *Ecological Informatics*, pages –.

Yen, G. G. and Fu, Q. (2002). Automatic frog call monitoring system: a machine learning approach. In *AeroSense 2002*, pages 188–199. International Society for Optics and Photonics.

Yuan, C. L. T. and Ramli, D. A. (2012). Frog sound identification system for frog species recognition. In *Context-Aware Systems and Applications*, pages 41–50. Springer.