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

August 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

Frogs play an important role in Earth's ecosystem, but the decline of their population has been spotted at many locations around the world. Monitoring frog activity can assist conservation efforts, and improve our understanding of their interactions with the environment and other organisms. Traditional observation methods require ecologists and volunteers to visit the field, which greatly limit the scale for acoustic data collection. Recent advances in acoustic sensors provide a novel method to survey vocalising animals such as frogs. Once sensors are successfully installed in the field, acoustic data can be automatically collected at large spatial and temporal scales. For each acoustic sensor, several gigabytes of compressed audio data can be generated per day, and thus large volumes of raw acoustic data are collected. To gain insights about frogs and the environment, classifying frog species in acoustic data is necessary. However, manual species identification is unfeasible due to the large amount of collected data, and enabling automated species classification has become very important. Previous studies on signal processing and machine learning for frog call classification often have two limitations: (1) the recordings used to train and test classifiers are trophy recordings (signal-to-noise ratio (SNR) ($\geq$ 15 dB); (2) each individual recording is assumed to contain only one frog species. However, field recordings typically have a low SNR ($<$ 15 dB) and contain multiple simultaneously vocalising frog species. This thesis aims to address two limitations, and makes the following contributions.

(1) Develop a combined feature set from temporal, perceptual, and cepstral domains for improving the state-of-the-art performance of frog call classification using trophy recordings (Chapter 3).

(2) Propose a novel cepstral feature via adaptive frequency scaled wavelet packet decomposition (WPD) to improve features' anti-noise ability for frog call classification using both trophy and field recordings (Chapter 4).

(3) Design a novel multiple-instance multiple-label (MIML) framework to classify multiple simultaneously vocalising frog species in field recordings (Chapter 5).

(4) Design a novel multiple-label (ML) framework to increase the robustness of classification results when classifying multiple simultaneously vocalising frog species in field recordings (Chapter 6).

Our proposed approaches achieves promising classification results compared with previous studies. With this developed classification techniques, the ecosystem at large spatial and temporal scales can be surveyed, which can help ecologists better understand the ecosystem.

Keywords

Acoustic event detection

Acoustic feature

Bioacoustics

Frog call classification

Multiple-instance multiple-label learning (MIML)

Multiple-label learning (ML)

Soundscape ecology

Syllable segmentation

Wavelet packet decomposition (WPD)

Acknowledgments

First, I would like to express my sincere gratitude and thanks to Dr. Jinglan Zhang (principal supervisor), for giving me an opportunity to study in Australia. During the entirety of this PhD study, I have learnt so much from her about having passion for work, combined with high motivation, which will benefit me throughout my life. I would also like to express my gratitude to Prof. Paul Roe (associate supervisor), for his consistent instructions and financial supports through the last three years.

I would also like to thank Dr. Michael Towsey (associate supervisor) for his provision of consistent guidance, discussions, and encouragement during my PhD study. Michael's attitude towards scientific research keeps motivating me go deeper into research.

I want to thank Prof. Vinod Chandran (associate supervisor) for his support in writing my confirmation report and this thesis. Vinod's strong background knowledge in signal processing greatly helps me improve my understanding of this research.

I would also like to express my gratitude to my family, especially my grandparents, parents and my wife. They have always supported my overseas study. Without their support, I could not give my full attention to PhD study and the completion of this thesis. My sincere thanks also go to all my friends for their love, attention and support to my PhD study.

Finally, I extend my thanks to the China Scholarship Council (CSC), Queensland University of Technology, and Wet Tropics Management Authority for their financial support.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AED	acoustic event detection
ANN	artificial neural network
AWSCCs	adaptive-frequency scaled wavelet packet decomposition sub-band cepstral coefficients
dB	decibel
DCT	discrete cosine transform
DFT	discrete Fourier transform
DT	decision tree
DTW	dynamic time warping
DWT	discrete wavelet transform
JCU	James Cook University
kNN	k-nearest neighbour
LDA	linear discriminant analysis
LPCs	linear predictive coefficients
MFCCs	Mel-frequency cepstral coefficients
MIML	multiple-instance multiple-label
ML	multiple-label
MLP	multiple layer perceptron

MWSCCs	Mel-frequency scaled wavelet packet decomposition sub-band cepstral coefficients
RBF	radial basis function
RF	random forest
SNR	signal-to-noise ratio
STFT	short-time Fourier transform
SVM	support vector machine
WPD	wavelet packet decomposition

Chapter 1

Introduction

1.1 Motivation

Frogs are greatly important for the Earth's ecosystem but their populations are rapidly declining. Frogs are an integral part of the food web and an excellent indicator for biodiversity due to their sensitivity to the environmental change [Böll et al., 2013]. Over the last two decades, rapid decline in frog populations has been spotted worldwide. This is regarded as one of the most critical danger to the global biodiversity. The causes for this decline are many, but global climate change [Carey and Alexander, 2003] and emerging diseases [Mutschmann, 2015] are thought as the biggest threats.

Developing techniques for monitoring frogs is becoming ever more important to gain insights about frogs and the environment. Since frogs employ vocalisations for most communications and have a small body size, they are often easier to be heard than seen in the field (Figure 1.1). This offers a possible way to study and evaluate frogs by detecting species-specific calls [Dorcas et al., 2009]. Duellman and Trueb [1994] classified frog vocalisations into six categories based on the context in which they occur: (1) mating calls, (2) territorial calls, (3) male release calls, (4) female release calls, (5) distress calls, and (6) warning calls. Among them, mating calls are now widely termed as advertisement calls. Most existing studies that using signal processing and machine learning to classify frog species use only advertisement calls for the experiment [Chen et al., 2012, Gingras and Fitch, 2013, Han et al., 2011, Huang et al., 2014, 2009]. This thesis will also use only advertisement calls for the experiment.

Traditional methods for classifying frog species, which require ecologists and volunteers

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

to physically visit sites, are costly and time-consuming. Although traditional methods can provide an accurate measure of daytime species richness, the scale limitation in both spatial and temporal domains is unavoidable. Recent advances in acoustic sensors provide a novel way to automatically survey vocal animals such as frogs. The use of acoustic sensors can greatly extend the spatial and temporal scales. Once acoustic sensors are successfully installed in the field, frog calls can be continuously collected. Each acoustic sensor can generate several gigabyte of compressed acoustic data, and so far large volumes of data has been collected and needs to be analysed. Consequently, enabling automated species classification in acoustic data has become increasingly important.

1.2 Research challenges

Most previous studies classify frog calls with trophy recordings, which are different from field recordings. Table 1.1 summarises the differences between trophy recordings and field recordings. Trophy recordings are collected in constrained environments with a directional microphone. In contrast, field recordings are collected in unconstrained environments with an omnidirectional microphone.

Based on these differences, two major challenges must be faced for building an accurate and robust frog call classification framework for field recordings:

1. Compared to trophy recordings which are collected in constrained environment with a directional microphone, field recordings tend to be noisy. Very often the desired signal

Table 1.1: Comparison between trophy and field recordings

Trophy recordings	**Field recordings**
Directional microphone	Omnidirectional microphone
High SNR for animals of interest ($\geq$ 15 dB) (Table A.1)	Low SNR for animals of interest (Table B.1)
One species per recording	Multiple species per recording
Close to animals	Far away from animals
Short recordings (seconds/minutes)	Long recordings (hours/days)

(frog call) is weak, and there are other overlapping signals such as bird calls and insect calls over frog calls. Therefore, features used for classifying frogs in field recordings must have a good anti-noise ability.

2. Most field recordings contain multiple frog species in an individual recording, which are different from recordings used in previous studies (one species per recording). The classification framework for studying frogs in field recordings must be able to classify multiple frog species for each individual recording.

1.3 Scope of PhD

The broad scope of this PhD research is to address the two aforementioned challenges, which could pave a way to successful classification of multiple simultaneously vocalising frog species in field recordings. The outcome of the research is of benefit to many applications of bioacoustics. Recordings used for the experiment are of two types: (1) trophy recordings, (2) field recordings. The use of trophy recordings allows our proposed methods to be easily compared to other published techniques. Successfully classifying frog species in field recordings can extend our proposed classification framework to address those recordings collected by acoustic sensors in real ecological investigations.

1.4 Original contributions

A frog call classification system often consists of three parts (Figure 1.2): (1)signal pre-processing, which includes signal processing, noise reduction, and syllable segmentation; (2) feature extraction (representing frog attributes into some feature vectors); and (3) classification (recognising

frog species using machine learning techniques).

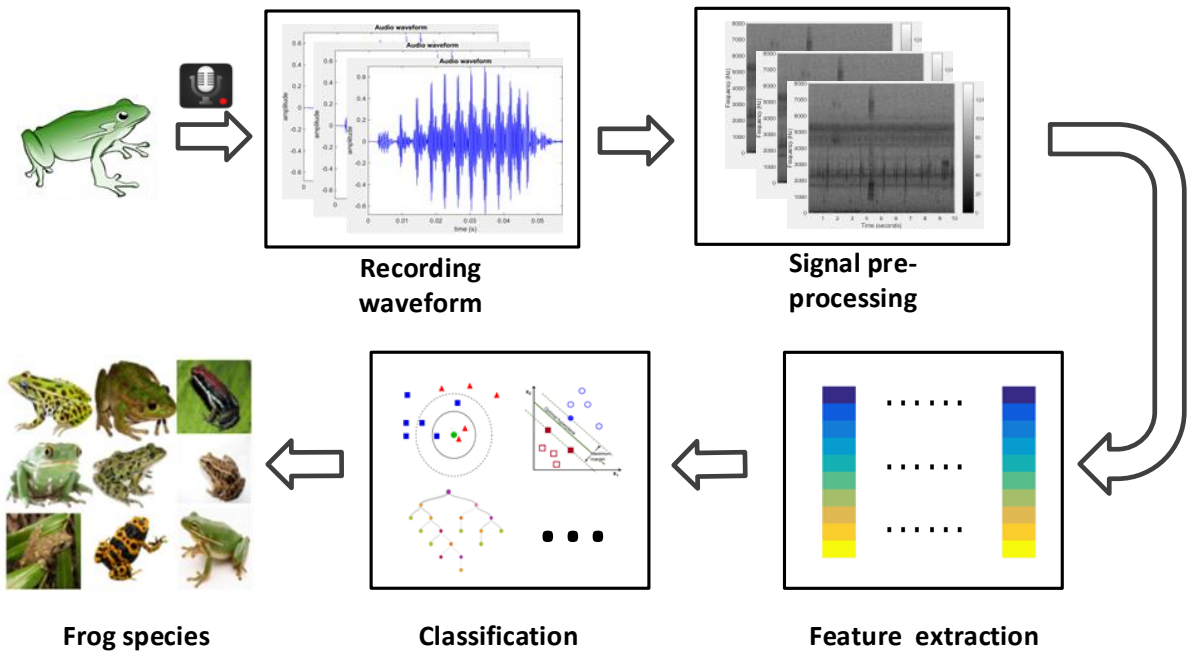


Figure 1.2: Flowchart of frog call classification: pre-processing, feature extraction, and classification

This research makes important contributions to the domains of syllable segmentation (one step in pre-processing), feature extraction, and classification. Specifically, this research proposes a novel acoustic event detection (AED) method to segment frog syllables in field recordings. To further improve the classification performance using trophy recordings, a combined feature set using temporal, perceptual, and cepstral features is constructed. To increase the anti-noise ability of cepstral features, a novel cepstral feature via adaptive frequency scaled wavelet packet decomposition (WPD) is developed. Moreover, two classification frameworks, multiple-instance multiple-label (MIML) classification and multiple-label (ML) classification, are adopted to cope with field recordings including multiple frog species. The detailed description of the contribution for each experiment is shown as follows:

1. Most previous studies test the proposed frog call classification methods using trophy recordings, and each individual recording is assumed to have only one frog species. The first experiment of this thesis aims to further improve the classification performance using trophy recordings. A novel feature combination using temporal, perceptual, and cepstral features is proposed for frog call classification. To reduce the bias of syllable

segmentation, Gaussian filtering is selectively used to remove the temporal gap within one syllable. Five feature sets are constructed using different combinations of temporal, perceptual, and cepstral features. Five machine learning algorithms are used for the classification. Experimental results on trophy recordings show that our proposed feature set outperforms other widely used feature sets for classifying frog calls.

This research has led to one ISSNIP conference paper, one Applied Acoustic journal article.

2. Since most field recordings are noisy, features' anti-noise ability is critical for achieving a good classification performance. The first experiment demonstrates that cepstral features used for classifying frog species in trophy recordings often have a high classification accuracy, but are very sensitive to the background noise. A novel cepstral feature is proposed via adaptive frequency scaled WPD for classifying frog species in both trophy and field recordings. Here, the adaptive frequency scale is generated by applying k-means clustering to the dominant frequencies of training dataset. Previous studies have shown that dominant frequencies of different frog species are different. A frequency scale, which fits the frequency distribution of different species, can increase the discriminability of cepstral features extracted by this scale. Experimental results show that our proposed cepstral feature not only achieves a higher classification accuracy but also has a better anti-noise ability.

   This research has led to one ICISP conference paper, one IEEE e-science conference paper, and one Ecological Informatics journal article.

3. Since most field recordings contain multiple simultaneously vocalising frog species per recording, the MIML classification framework is a natural fit for those frog recordings, and enables the classification of multiple simultaneously vocalising frog species. A novel AED method is proposed for the segmentation of frog syllables in field recordings with limited annotated data. Compared to other AED methods, our proposed AED method can achieve the best syllable segmentation results, which is verified by the MIML classification results. For each segmented frog syllable, event based features are calculated. Three MIML classifiers are used for the classification with three feature sets. Experimental results demonstrate that MIML learning can classify multiple frog species in field recordings with a hamming loss of 0.1192, which is 2.7 times better than the non-informative

classifier. Experimental results also show that the proposed MIML classification framework can achieve better performance compared to the SISL classification.
This research has led to one ICISP conference paper.

4. For the MIML classification, the results are highly affected by the AED results. To further improve the classification performance, one solution is to prepare large volumes of annotated acoustic data and apply supervised learning algorithms for improving segmentation results. Another is to use a different framework without the need of syllable segmentation. This thesis examines the latter option and adopts ML learning to classify multiple simultaneously vocalising frog species in field recordings. Three global features are first extracted from each individual recordings: linear prediction coefficients (LPCs), Mel-frequency cepstral coefficients (MFCCs), and adaptive-frequency scaled wavelet packet decomposition sub-band cepstral coefficients (AWSCCs). Two cepstral features are constructed using statistical analysis and spectral clustering. A novel feature set of LPCs and AWSCCs is used for the ML classification. Experimental results show that ML classification can achieve similar performance with MIML classification.

   This research has led to a ICCS conference paper.

1.5 Associated publications

Below is a list of the publications arising from this PhD research:

Journal Articles

1. **Xie, Jie**, Towsey, Michael, Zhang, Jinglan, and Roe, Paul, Frog call classification based on enhanced features and machine learning algorithms, Applied Acoustics, Volume 113, June 2016, pp. 193-201.

   This work corresponds to Chapter 3 in this thesis, which presents a combined feature set for frog call classification in trophy recordings.

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

This work corresponds to Chapter 4 in this thesis, which develops a novel cepstral feature for frog call classification in both trophy and field recordings.

3. 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.

4. **Xie, Jie**, Towsey, Michael, Zhang, Jinglan, and Roe, Paul, Frog call classification: a survey, Artificial Intelligence Review (Accepted)

   This work corresponds to Chapter 2 in this thesis, which reviewed the extant literature on frog call classification.

5. **Xie, Jie**, Towsey, Michael, Zhang, Jinglan, and Roe, Paul, Classification of Frog Vocalizations using Acoustic and Visual Features, Journal of Signal Processing Systems (Under review with minor revision)

Conference Papers

1. **Xie, Jie**, Michael Towsey, Jinglan Zhang, Paul Roe, Detecting Frog Calling Activity Based on Acoustic Event Detection and Multi-label Learning, Procedia Computer Science, Volume 80, 2016, Pages 627-638.

   This work corresponds to Chapter 5 in this thesis, which applied ML learning for frog call classification.

2. **Xie, Jie**, Towsey, Michael, Zhang, Liang, Yasumiba, Kiyomi and Schwarzkopf, Lin, 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. Springer International Publishing, 2016, pp 222-230.

   This work corresponds to Chapter 6 in this thesis, which applies MIML learning for frog call classification.

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, Springer International Publishing, 2016, pp 231-239.

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.

1.6 Thesis structure

This thesis is organised in the manner outlined as follows:

Chapter 1 provides a brief introduction to the problem of "Frog call classification using machine learning algorithms". The ecological significance of studying frogs is first illustrated. Then, two methods for frog monitoring are compared, and two challenges are identified. In the following chapters, we will see that the methods proposed in this thesis are driven by the motivation of solving those two challenges.

Chapter 2 reviews the significant and latest literature of frog call classification using machine learning techniques. Three main parts of a frog call classification framework are discussed: signal pre-processing, feature extraction, and classification. In addition, evaluation metrics and previous experimental results are presented. This chapter provides a foundation for the research problem and necessary information about the state-of-the-art frog call classification methods. Meanwhile, the research gap is identified, which points out the potential research direction.

Chapter 3 develops a combined feature set for frog call classification using trophy recordings. A combination of temporal, perceptual, and cepstral features is used for frog call classification. Classification results of five machine learning algorithms are compared to our combined feature set.

Chapter 4 investigates WPD for extracting a novel cepstral feature. An adaptive frequency scale is first generated by applying k-means clustering to dominant frequencies of those frog species to be classified. Then, *adaptive frequency scaled WPD* is used for calculating a novel cepstral feature. Two machine learning algorithms are used for the classification. The propose cepstral feature will be used in Chapter 6 as well.

Chapter 5 discusses the limitations of traditional SISL classification framework for classifying multiple simultaneously vocalising frog species in field recordings, and adopts the MIML classification framework to classify frog species in those recordings. A novel AED method is developed for frog syllable segmentation. Various event based features are extracted from each individual syllable. A bag generator is used for constructing a bag-level feature. Finally, three MIML classifiers are used for the classification.

Chapter 6 investigates the shortcomings of the MIML classification framework, and introduces ML learning for classifying multiple frog species in field recordings. Three global features are calculated without the segmentation process: LPCs, MFCCs, and AWSCCs. Two cepstral-feature sets are constructed using statistical analysis and spectral clustering. Three ML classifiers are used for the classification with constructed feature sets.

Chapter 7 summarises the major achievements of this thesis and analyses the limitations of developed approaches. Some directions of future work are also pointed out.

Chapter 2

An overview of frog call classification

This chapter reviews the extant literature on frog call classification using machine learning algorithms. To the best of this author's knowledge, no previous studies focus on frog call classification using multiple-instance multiple-label (MIML) or multiple-label (ML) learning. Therefore, this chapter will mainly review the single-instance single-label (SISL) learning for frog call classification. For MIML and ML learning, some prior work on bird call classification is reviewed. This review mainly aims to give a quantitative and detailed analysis of related techniques for frog call classification. Then, several major challenges that have not been addressed in prior work are identified, and hence the advances in this thesis are necessary and significant. Detailed information of each part will be described in following sub-section.

2.1 Overview

Three parts play important roles in the performance of frog call classification: signal pre-processing, feature extraction, and classification. Figure 1.2 depicts the common structure of frog call classification.

2.2 Signal pre-processing

Signal pre-processing contains signal processing, noise reduction, and syllable segmentation.

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). Techniques used for frog signal processing include STFT [Colonna et al., 2015, Huang et al., 2014, 2009], WPD [Yen and Fu, 2002], and DWT [Colonna et al., 2012b]. STFT is the most widely used technique due to its flexible implementation and better applicability. Given one frog call $x(n)$, 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 Figure 2.1.

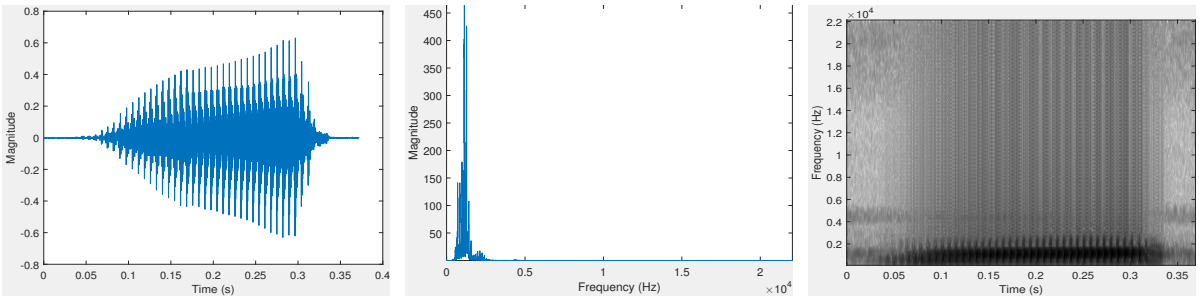


Figure 2.1: Waveform, spectrum and spectrogram of one frog syllable for *Mixophyes fasciolatus*. The window function, size and overlap are Hamming window, 128 samples and 85%, respectively

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. A wavelet threshold function in the one-dimensional signal was used as the filter kernel function. Bedoya et al. [2014] introduced a spectral noise gating method for noise reduction. Specifically, the selected frequency band spectrum of the frogs' call to be detected was estimated and suppressed. Although the aforementioned noise reduction methods can reduce the background noise, some of the desired signals will be suppressed. Noise reduction are thus selectively used based on the SNR of 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 accuracy of syllable segmentation will directly affect the classification performance, because features for frog call classification are calculated from each segmented syllable. Frog syllable segmentation methods in previous studies are summarised and listed in Table 2.1. However, all previous methods cannot address recordings with multiple simultaneously vocalising frog species. Meanwhile, those methods, which use temporal features for segmentation, cannot address field recordings.

Table 2.1: Summary of prior work for frog syllable segmentation. Here, *E* denotes energy, *ZCR* denotes zero-crossing rate. Sequential denotes that syllables are segmented using the same sequence as those syllables in the recording

Authors	Features for segmentation	Procedure
Han et al. [2011]	Spectral entropy	Manual
Jaafar et al. [2013b]	E and ZCR	Sequential
Huang et al. [2009]	Amplitude	Non-sequential
Härmä [2003]	Spectrogram	Non-sequential
Colonna et al. [2015]	Incremental E and Incremental ZCR	Sequential and real time

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 a high classification performance [Fox, 2008]. For frog call classification, acoustic features can be classified into five categories: temporal features, perceptual features, time-frequency features, cepstral features, and other features.

2.3.1 Temporal and perceptual features for frog call classification

Temporal features for frog call classification have been explored for a long time [Camacho et al., 2013, Chen et al., 2012, Dayou et al., 2011, Huang et al., 2014, 2009, 2008]. To achieve a better classification performance, temporal features are often combined with perceptual features for frog call classification.

Huang et al. [2009] used spectral centroid, signal bandwidth, and threshold-crossing rate for frog call classification with kNN and 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 ANN. 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: kNN, a multivariate Gaussian distribution model and 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. [2013] 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, one-dimensional recording waveform is often transformed into its two-dimensional time-frequency representation. Then, features based on the time-frequency representation are computed 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: LDA, DT, and SVM. Brandes [2008] proposed a method for classifying animal calls using duration, maximum frequency, and frequency bandwidth, and with HMM used as the classifier. Yen and Fu [2002] combined wavelet transform and two different dimensionality reduction algorithms to produce the final feature. Then, a NN classifier is used for frog call classification. Grigg et al. [1996] developed a system to monitor the effect of the introduced Cane Toad on the frog population of Queensland. The

[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.

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 novel feature set for detecting frogs with a similarity measure based on Euclidean distance. The feature set contained dominant frequency, frequency difference between the lowest and dominant 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.

2.3.3 Cepstral features for frog call classification

Cepstral features (MFCCs) are popular for frog call classification. Jaafar et al. [2013a] introduced MFCCs and LPCs as features. Then kNN and SVM were used as classifiers for frog call identification. Yuan and Ramli [2013] also used MFCCs and LPCs as features. Then kNN was used 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 LAMDA for frog call recognition. Vaca-Castano and Rodriguez [2010] proposed a method to identify animal species, which consisted of MFCCs, PCA and kNN. Jaafar et al. [2013b], Tan et al. [2014] published three papers about frog call classification using MFCCs, Δ MFCC and $\Delta\Delta$ MFCC calculated as features. Then kNN and SVM were used for classification. Colonna et al. [2012a] introduced MFCCs for classifying anurans with kNN.

2.3.4 Other features for frog call classification

Besides temporal features, perceptual features, time-frequency features, and cepstral features, other features are 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 the vocalisation waveform envelope as features, then classified calls by matching the extracted envelope with the original signal envelope. Kular et al. [2015] treated the sound signal of a frog call as a texture image. Then, texture visual words and MFCCs were calculated for frog call classification.

2.4 Classifiers

For frog call classification, numerous pattern recognition methods have been used to construct the classifier, such as Bayesian classifier [Brandes et al., 2006], kNN [Colonna et al., 2012a, Dayou et al., 2011, Gingras and Fitch, 2013, Han et al., 2011, Huang et al., 2009, 2008, Jaafar et al., 2013a,b,b, Vaca-Castano and Rodriguez, 2010, Yuan and Ramli, 2013], SVM [Acevedo et al., 2009, Gingras and Fitch, 2013, Huang et al., 2009, 2008, Jaafar et al., 2013a, Tan et al., 2014], HMM [Brandes, 2008], GMM [Gingras and Fitch, 2013, Huang et al., 2008], NN [Huang et al., 2014, Yen and Fu, 2002], DT [Acevedo et al., 2009, Grigg et al., 1996], one-way multivariate ANOVA [Camacho et al., 2013], and LDA [Acevedo et al., 2009, Lee et al., 2006]. Besides classifiers, other methods for classifying frog species included 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]. The summary of classifiers for frog call classification is listed in Table 2.2. kNN is the most commonly used classifier for its simplicity and easy application. However, kNN is sensitive to the local structure of the data, as well as to the distance and distance function. Therefore, kNN is often run multiple times based on different initial points. SVM is another widely used classifier for its good generalisation ability. However, the performance of SVM is quite sensitive to the selection of the regularisation and kernel parameters, and it is possible to over-fit when tuning these hyper-parameters. Since selecting suitable parameters for SVM is very important, most previous studies conducted the parameter setting by grid search [Hsu et al., 2003].

2.5 MIML or ML learning for bioacoustic signal classification

To the best of this author's knowledge, there is still no paper that uses MIML or ML learning to focus on frog call classification. In contrast, some previous research has applied MIML or ML learning to study bird calls.

For MIML learning, Briggs et al. [2012] introduced the MIML classifiers for acoustic classification of multiple simultaneously vocalising bird species. In their method, a supervised learning classifier (random forest) was first employed for segmenting acoustic events. Then features were extracted from each segmented acoustic event. Before putting features into

Table 2.2: A brief summary of classifiers in the literature

Reference	Classifier	Reference	Classifier
Brandes et al. [2006]	Bayesian classifier	Acevedo et al. [2009]	Support vector machine
Huang et al. [2008]	K-nearest neighbour	Huang et al. [2009]	Support vector machine
Huang et al. [2009]	K-nearest neighbour	Tan et al. [2014]	Support vector machine
Vaca-Castano and Rodriguez [2010]	K-nearest neighbour	Gingras and Fitch [2013]	Support vector machine
Dayou et al. [2011]	K-nearest neighbour	Jaafar et al. [2013a]	Support vector machine
Han et al. [2011]	K-nearest neighbour	Xie et al. [2015c]	Support vector machine
Gingras and Fitch [2013]	K-nearest neighbour	Brandes [2008]	Hidden Markov model
Jaafar et al. [2013b]	K-nearest neighbour	Huang et al. [2008]	Gaussian mixture model
Jaafar et al. [2013a]	K-nearest neighbour	Gingras and Fitch [2013]	Gaussian mixture model
Yuan and Ramli [2013]	K-nearest neighbour	Huang et al. [2014]	Neural networks
Jaafar and Ramli [2013]	K-nearest neighbour	Yen and Fu [2002]	Neural networks
Xie et al. [2015b]	K-nearest neighbour	Grigg et al. [1996]	Decision tree
Xie et al. [2015d]	K-nearest neighbour	Acevedo et al. [2009]	Decision tree
Xie et al. [2015a]	K-nearest neighbour	Camacho et al. [2013]	One-way multivariate ANOVA
Colonna et al. [2012a]	K-nearest neighbour	Acevedo et al. [2009]	Linear discriminant analysis
Huang et al. [2008]	Support vector machine	Lee et al. [2006]	Linear discriminant analysis

classifiers, a bag generator was used to construct a bag-level feature. Lastly, three MIML classifiers were experimentally evaluated: MIML-SVM, MIML-RBF, and MIML-kNN. Dufour et al. [2013b] used MFCCs and three MIML classifiers to classify birds. To be specific, MFCCs were first calculated for each frame. Then two new feature vectors were computed to represent longer segments. Lastly, three MIML classifiers were experimentally evaluated: MIML-RBF, MIML-kNN, and M3MIML (Maximum Margin Method for Multi-instance Multi-label Learning).

For ML learning, several papers have been published in the Neural Information Processing Scaled for Bioacoustics (NIPS4B challenge) [Glotin et al., 2013a], which classified birds, insects, and amphibians recordings, followed by signal pre-processing, segmentation, feature extraction, feature selection, and classification [Chen et al., 2013, Lasseck, 2013, Massaron, 2013, Mencıa et al., 2013, Stowell and Plumbley, 2013]. Lasseck [2013] first processed the recordings via the application of STFT, noise reduction and segmentation. Then, file-statistics, segment-statistics and segment-probabilities were calculated as features. Finally, an ensemble of randomised decision trees was used for the classification of each sound class. Stowell and Plumbley [2013] used either MFCC statistics (52 dimensions), chirplet histograms (up to 20,000 dimensions), or both, as features. Then, random forest was used for ML classification. Mencıa et al. [2013] proposed a new feature extraction method via an unsupervised generation of an aleatory number of features, which included random patching, a de-noising auto-encoder unit and subsequent convolution. For the classification, a pairwise ensemble of SVMs, random decision trees and a single layer NN were used. Massaron [2013] described an approach

that involved building an ensemble of generalised linear models and a classification model by hinge loss and program based on stochastic gradient descent optimisation, and boosted trees ensembles. Chen et al. [2013] first calculated prominent features from windowed MFCCs, and leveraged them to build an ensemble classifier which was a blend of different classifiers (Gradient Boosting Tree models, random forest, Lasso and elastic-net regularized generalised linear model).

2.6 Experiment results of state-of-the-art frog call classification

2.6.1 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 recognised positives; (2) TN: correctly recognised negatives; (3) FN: positives recognised as negatives; (4) FP: negatives recognised as positives. Then, accuracy, precision, recall (sensitivity), specificity, and false positive rate can be defined as

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

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

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

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

$$FPR = \frac{FP}{FP + TN} \tag{2.6}$$

2.6.2 Previous experimental results

Table 2.3 shows the list of summarised frog call classification methods, together with the database they used and corresponding performance. From Table 2.3, we can find that few studies explored frog vocalisations using signal processing and machine learning techniques before 2010. Due to the decrease of frog biodiversity, advances in signal processing, machine learning techniques and acoustic sensors, the research in frogs has been increased in the last five years. In addition, few datasets are publicly shared with researchers. The classification performance of different studies varies a lot. One of the main reasons is the use of different datasets.

2.7 Summary of research gaps

This chapter reviews three main parts of a frog call classification system: signal pre-processing, feature extraction, and classification. The review also points out the research gaps in current literature as follows:

2.7.1 Database

One major problem for frog call classification is the lack of a universal database. The databases used in prior work are often related to geographical regions, because researchers from different countries focus on particular frog species in their specific area (Table 2.3). Therefore, it is difficult for researchers to compare their particular classification frameworks. More data is needed to compare the performance and to improve the robustness of the classification frameworks.

2.7.2 Signal pre-processing

Currently, 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.

Noise reduction is an optional processing step for frog call classification. For some databases used as shown in Table 2.3, frog recordings have a high SNR, where noise reduction is unnecessary. However, when studying field recordings, noise reduction is essential for improving the

Table 2.3: A brief overview of frog call classification performance

Database	Performance	Reference	Data source
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
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)
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)
3 frog species with 50 samples	Averaged classification accuracy of 90%	Dang et al. [2008]	Unknown
5 frog species with 727 syllables	Averaged classification accuracy of 95.86%	Huang et al. [2008]	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
5 frog species with 959 samples	Averaged classification accuracy of 90.03%	Huang et al. [2009]	Unknown
12 frog species with 379 samples, 10 bird species with 193 samples	Averaged classification accuracy of 86.6%	Vaca-Castano and Rodriguez [2010]*	Recorded in Puerto Rico (http://www.amazon.com/Los -Anfibios-Reptiles-Puerto-Rico/dp/084770243X) (available)
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
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)
9 frog species with 49 samples	Averaged classification accuracy of 97.60%	Colonna et al. [2012a]	Collected on the campus of the Federal University of Amazonas in Manaus, Brazil (unavailable)
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)
3 frog species with 635 calls	Precision of 99%, recall of 92%	Camacho et al. [2013]	Collected from Costa Rica (unavailable)
142 species belonging to four genera	Genus classification accuracy above 70%	Gingras and Fitch [2013]	obtained from commercially available compact discs (CDs) (available)
8 frog species with 160 samples	averaged classification accuracy of 98.1%	Yuan and Ramli [2013]	Obtained from AmphibiaWeb (http://amphibiaweb.org/)(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)
10 frog species with 250 syllables	Averaged classification accuracy of 98.8%	Jaafar et al. [2013a]	Internet database (http://learning.froghome.org/) and IBM,USM (http://www.frogwatch.org.au/?action=animal.list) (available)
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)
13 frog species with 1514 samples	Averaged recognition rate of 93.4%	Huang et al. [2014]	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
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)
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)

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 field

recordings which are very noisy. This method cannot segment frog syllables in field recordings with multiple frog species. Recent use of unsupervised learning algorithms opens a path for segmenting frog syllables in field recordings with multiple frog species. However, like other unsupervised algorithms, this method has a disadvantage that not all segmented syllables are frog vocalisations [Potamitis, 2015]. Briggs et al. [2012] used a supervised random forest for bird call segmentation. However, this method required lots of tagged acoustic data to train the classifier.

For syllable segmentation, temporal features are more sensitive to the background noise than perceptual features, because different frequency components can be separated by transforming the signal from temporal domain to perceptual domain. However, temporal features cannot segment frog syllables in field recordings with multiple frog species, because temporal features have no ability to separate different frequency components. Compared to temporal features, the use of amplitude-frequency information provides a robust method to segment field recordings.

2.7.3 Acoustic features

Most previous studies directly transplant features developed for speech recognition to analyse frog calls, which might not be suitable. For example, MFCCs, which are based on the calculation of a non-linear Mel-scale filter-bank, are designed for studying speech. The Mel-scale is designed for the perceptual scale of pitches judged by human listeners. The frequency distribution might not be suitable for studying frogs. The direct use of speech features will therefore restrict classification performance.

Most perceptual features are extracted by directly calculating the statistics over frames, which will loss the temporal information. To add the temporal information of the feature set, temporal features can be combined with perceptual features and cepstral features to achieve higher classification accuracy. Transforming audio data into its 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. Image features derived from spectrograms are worth investigating for frog call classification. Sparse coding has been widely applied for feature extraction in other scaled bioacoustics studies [Glotin et al., 2013b, Razik et al., 2015], which could be a potential direction for frog call classification.

2.7.4 Classifiers

Almost all previous studies assume that each recording has only one frog species, and then a SISL 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 a frog chorus (Figure 2.2). Based on this attribute of frog call recordings, the classification problem can be naturally framed as a MIML classification or a ML classification problem rather than SISL classification. In previous studies, MIML and ML learning have been used to solve bioacoustic problems, but mainly focus on birds.

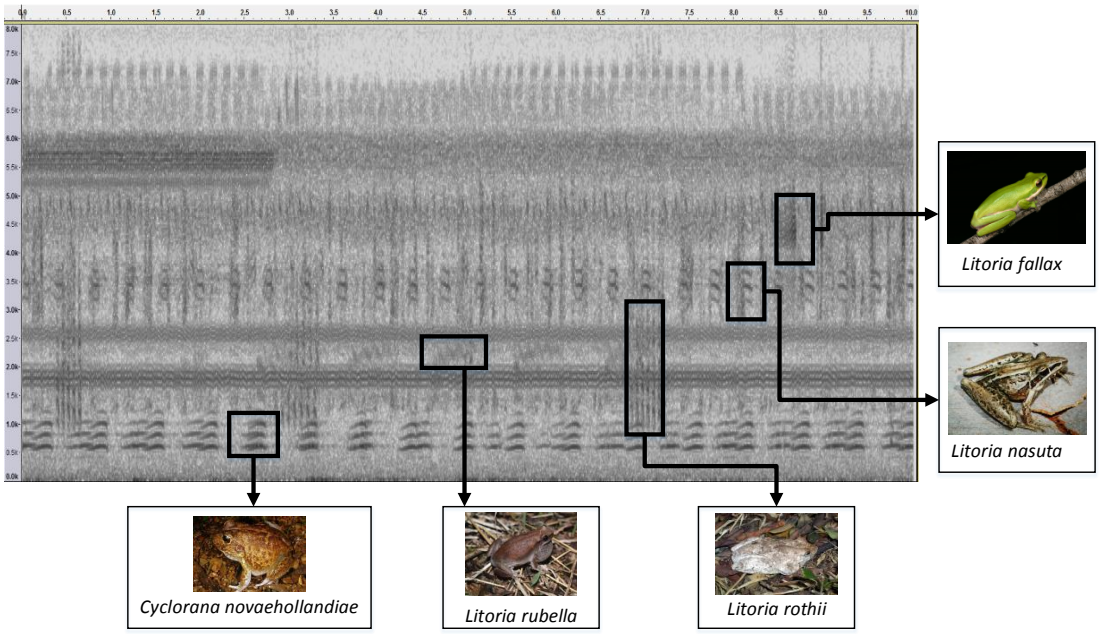


Figure 2.2: An example of field recording with multiple simultaneously vocalising frog species. Five frog species exist in this 10-second recording: *Cyclorana novaehollandiae*, *Litoria rubella*, *Litoria nasuta*, *Litoria rothii*, and *Litoria fallax*

In the following chapters, we mainly aim to address the above identified research gaps in signal pre-processing, feature extraction, and classification. To do so, four experiments are designed and organised in the manner outlined in Figure 2.3.

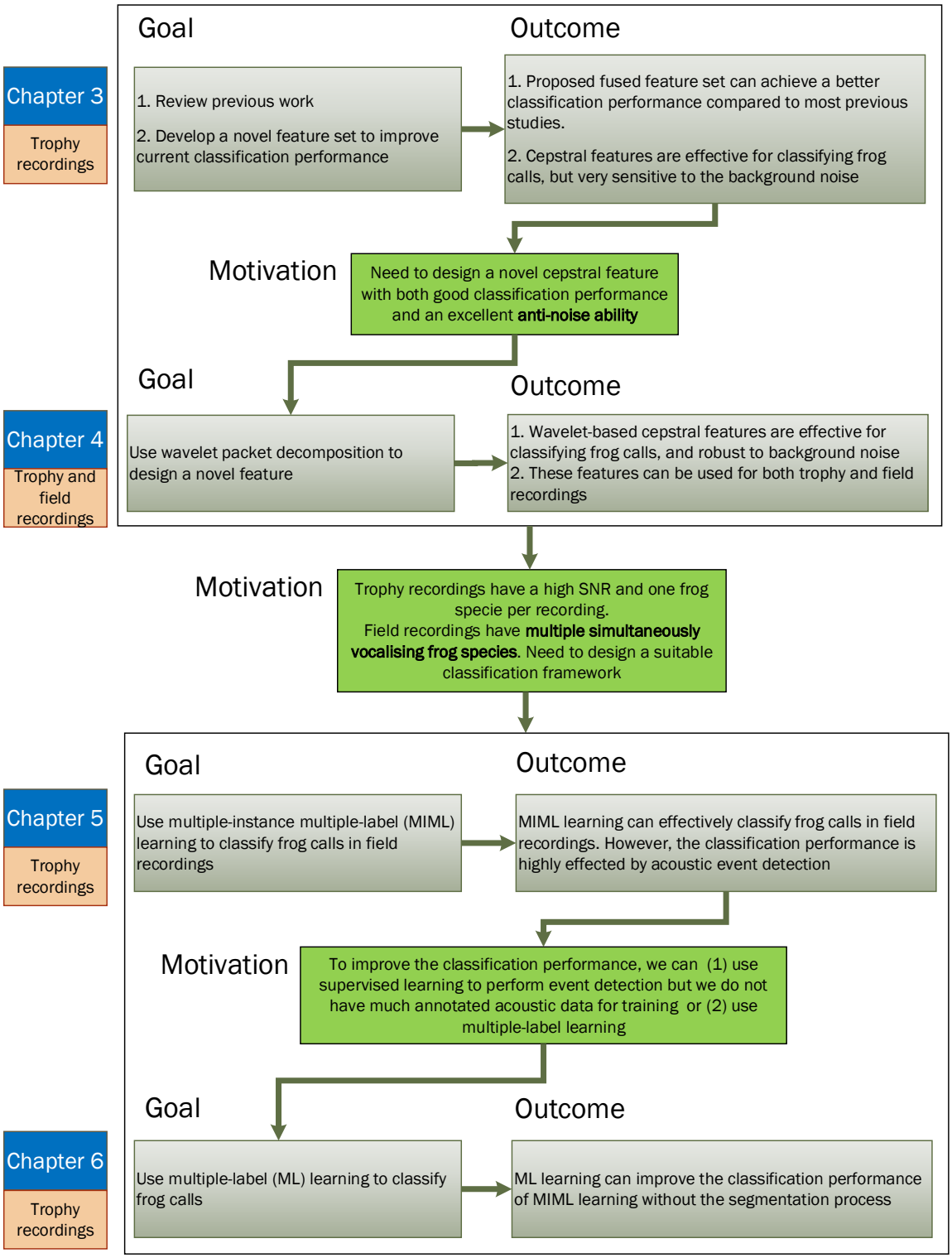


Figure 2.3: Structure of the four main chapters of this thesis

Chapter 3

Frog call classification based on feature combination and machine learning algorithms

---

Research problem

Previous studies classified frog calls in trophy recordings using one or two types of features from temporal, perceptual, and cepstral features. However, a combined feature set using one or two types of features cannot classify frog species that share similar characteristics in one or two domains (temporal domain, perceptual domain, and cepstral domain).

Research sub-question

How to build a feature set for further improving frog call classification performance in trophy recordings?

3.1 Overview

This chapter aims to compare various feature sets using different machine learning algorithms, and finds the best feature set for classifying frogs in trophy recordings. Based on the classification performance, suggested features can be adapted to study field recordings. In particular, we want to know which feature can be adopted from classifying frog species in trophy recordings to field recordings, because the final goal of this thesis is to classify multiple simultaneously vocalising frog species in field recordings.

The proposed method is evaluated using twenty-four frog species, which are geographically well distributed through Queensland, Australia. Five feature sets are constructed and evaluated

using five machine learning algorithms.

3.2 Methods

Our frog call classification system contains six modules (Figure 3.1): data description, syllable segmentation, pre-processing, feature extraction, feature combination, and classification. Detailed information of each module is described in following subsections. Different from [Huang et al., 2009], pre-processing is directly applied to segmented syllables rather than continuous recordings.

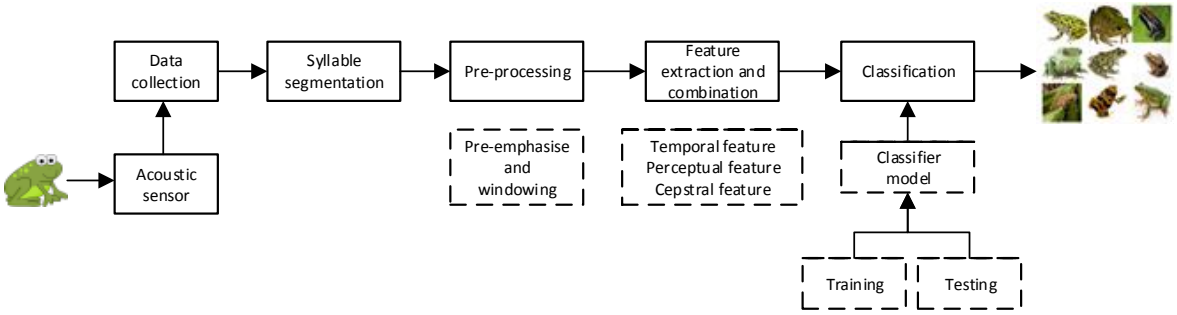


Figure 3.1: Flowchart of frog call classification system using the combined feature set

3.2.1 Data description

In this chapter, twenty-four frog species, which are widespread in Queensland, Australia, are selected for experiments (Table 3.1). All those frog species are collected from trophy recordings [Stewart, 1999]. All the recordings are two-channel, sampled at 44.10 kHz and saved in MP3 format. Similar with other trophy recordings, used recordings are obtained with a directional microphone and have a high SNR. Each recording includes one frog species with the duration ranging from eight to fifty-five seconds.

3.2.2 Syllable segmentation based on an adaptive end point detection

Each recording is made up of multiple continuous calls of one frog species. For frogs, one syllable is an elementary acoustic unit for classification, which is a continuous vocalisation emitted from an individual [Huang et al., 2009]. In this chapter, one approach built on the H$\ddot{a}$rm$\ddot{a}$'s method is used to perform syllable segmentation for frog calls [Härmä, 2003]. The syllable

Table 3.1: Summary of scientific name, common name, and corresponding code. Frog species name with asterisk means that it needs to be smoothed before segmentation

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

segmentation process is based on the spectrogram, which is generated by applying STFT to each recording waveform. 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 Figure 3.2, which is based on the iterative frequency-amplitude information of the spectrogram. This chapter focuses on the evaluation of combined 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 is smaller than 15% of the maximum energy and larger than 1.5 times the averaged energy are removed for each frog species experimentally [Gingras and Fitch, 2013].

In this chapter, smoothing spectrogram is optionally applied to the spectrogram before the H*ä*rm*ä*'s algorithm, because some frog species have a large temporal gap within one syllable

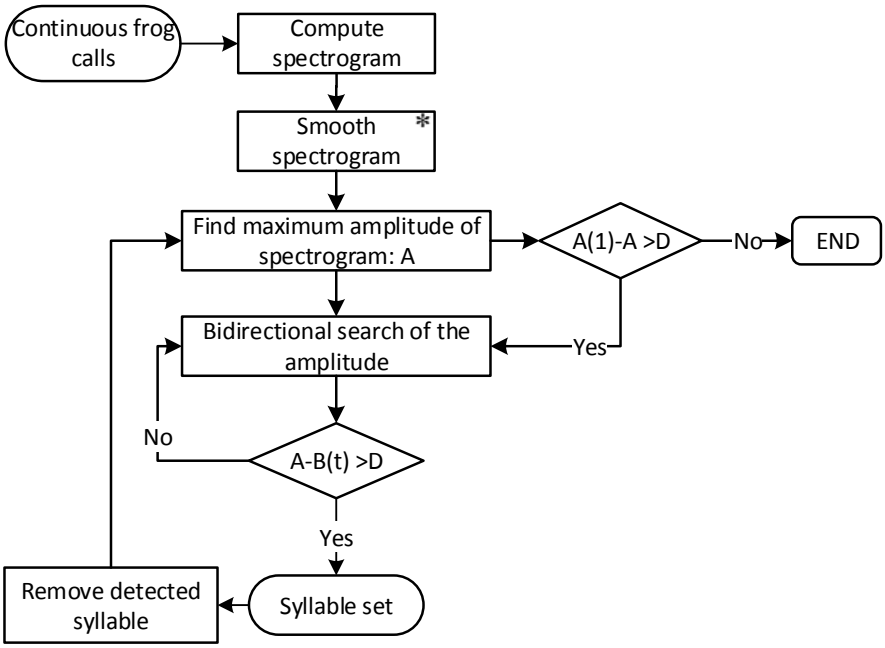
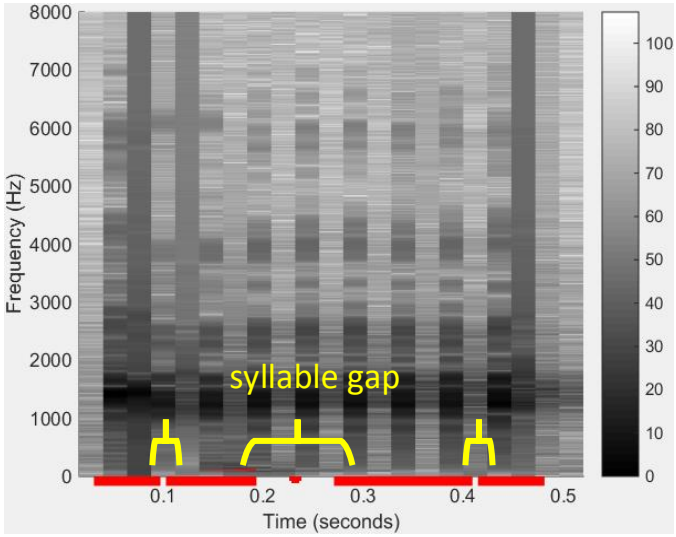


Figure 3.2: Segmentation method based on Härmä's algorithm. Here, D is the amplitude threshold for stopping criteria which is set at 20 dB experimentally, and the segmentation result is sensitive with this value. A is the maximum amplitude value of the spectrogram and we save the first maximum amplitude as $A(1)$, $B(t)$ is the amplitude of frame t. An asterisk denotes the optional processing step

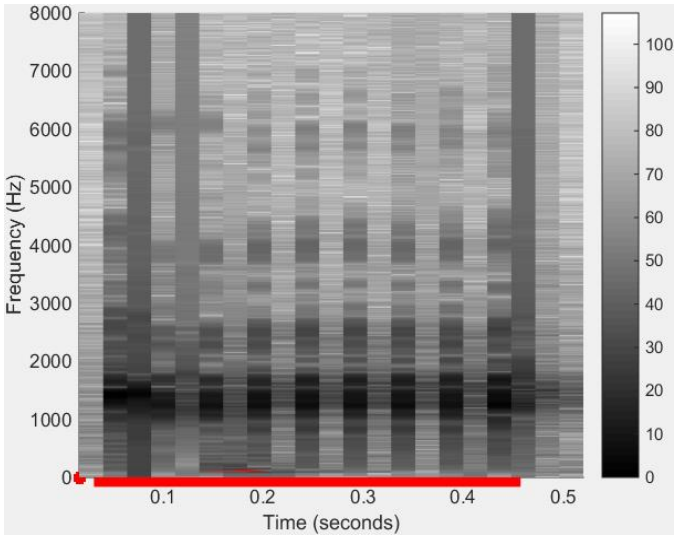
(see Figure 3.3). As for the smoothing, a Gaussian filter (7×7) is applied to the spectrogram, where the size is set by taking into account a trade-off between connecting gaps within one syllable and separating adjacent syllables. The segmentation result after smoothing is shown in Figure 3.3. The distribution of number of syllable for all frog species after segmentation is shown in Figure 3.4.

3.2.3 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:



(a) Before smoothing



(b) After smoothing

Figure 3.3: Syllable segmentation results are marked with a red line for *Neobatrachus sudelli* (one syllable)

Pre-emphasise

Some collected frog calls have low amplitude but in the high frequency, which will have an effect on feature extraction of the spectrum at the high frequency end. To enhance those high-frequency components and reduce the low-frequency components, a first-order high-pass filter with finite impulse response is introduced and defined as

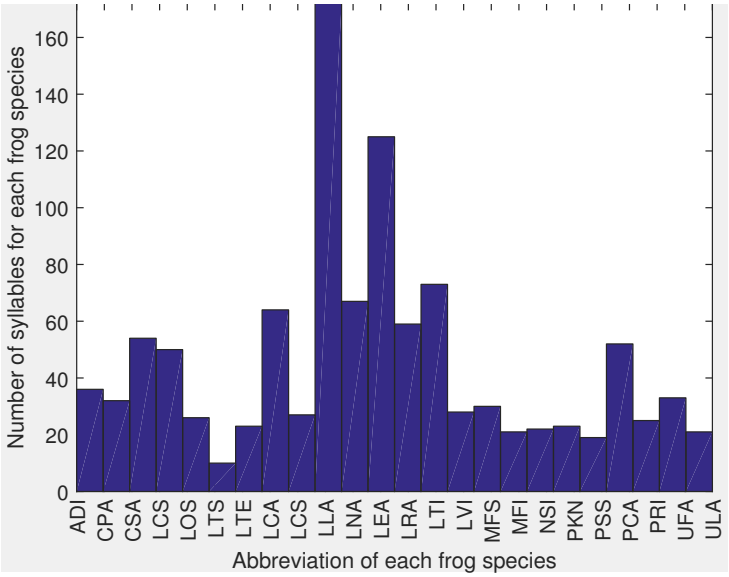


Figure 3.4: Distribution of number of syllable for all frog species. The x-axis is the abbreviation of each frog species, and the corresponding scientific name can be found in Table 3.1

$$y(n) = s(n) - \alpha s(n-1) \tag{3.1}$$

where s denotes a frog syllable, y 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.

Windowing

After pre-emphasis, 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

$$w(n) = 0.54 - 0.46cos(\frac{2n\pi}{L-1}), 0 \leq n \leq L-1 \tag{3.2}$$

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 chapter. The signal after windowing process is expressed as

$$x(n) = w(n)y(n) \tag{3.3}$$

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 LPCs and MFCCs [Bedoya et al., 2014, Jaafar and Ramli, 2013, Yuan and Ramli, 2013]. MFCCs achieved better performance than LPCs [Yuan and Ramli, 2013]. Different from the hybrid feature sets using in [Gingras and Fitch, 2013, Han et al., 2011, Huang et al., 2009], our combined feature set consists of more features, such as oscillation rate [Xie et al., 2015b], to further improve the classification accuracy. In this chapter, temporal features include syllable duration, Shannon entropy, rényi entropy, zero-crossing rate, averaged energy, and oscillation rate. Perceptual features contain spectral centroid, spectral flatness, spectral roll-off, signal bandwidth, spectral flux, and fundamental frequency. Here the word *perceptual* is defined according to [Lei et al., 2014]. MFCCs are used as a cepstral feature. The description of each feature is listed below.

(1) Syllable duration: Syllable duration [Xie et al., 2015b] is directly obtained from the bounds (time domain) of the segmentation results.

$$Dr = x(n_e) - x(n_s) \tag{3.4}$$

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

(2) Shannon entropy: 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.

$$Se = -\sum_{i=1}^{L} p_i log_2(p_i) \tag{3.5}$$

where L is the length of a frog syllable.

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

$$Re = \frac{1}{1-\alpha} log_2(\sum_{i}^{n} p_i^{\alpha}) \tag{3.6}$$

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

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

$$Zcr = \frac{1}{2}\sum_{n=0}^{L-1}[sgn(x(n)) - sgn(x(n+1))] \tag{3.7}$$

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

$$Ae = \frac{1}{L}\sum_{n=0}^{L-1} x(n)^2 \tag{3.8}$$

(5) Oscillation rate: 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 are discarded due to the uncertainty. Next, the autocorrelation is performed by the length of the vector. Furthermore, a DCT is employed to the vector after mean subtraction, and the position of the highest frequency is achieved to calculate the oscillation rate.

(6) Spectral centroid: 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.

$$Sc = \frac{\sum_{k=0}^{N-1} f_k X(k)}{\sum_{k=0}^{N-1} X(k)} \tag{3.9}$$

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

of the power spectrum and defined as

$$Sf = \frac{exp(\frac{1}{N}\sum_{k=0}^{N-1} InX(k))}{\frac{1}{N}\sum_{k=0}^{N-1} X(k)} \tag{3.10}$$

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

$$\sum_{k}^{H} X(k) = \theta \sum_{k=1}^{N-1} X(k) \tag{3.11}$$

where θ is set at 0.85, experimentally.

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

$$Bw = \sqrt{\frac{\sum_{k=0}^{N-1}(k - Sc)^2|x(n)|}{\sum_{k=0}^{N-1} X(k)}} \tag{3.12}$$

(10) Spectral flux: 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 denoted as

$$Sf = \sum_{k=-\frac{N}{2}}^{k=\frac{N}{2}} H[|X(n,k)| - |X(n-1,k)|] \tag{3.13}$$

where $H(x) = (x + |x|)/2$ is a 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 chosen or specified threshold, the frequency of that peak will be selected to calculate the fundamental frequency.

(12) Mel-frequency cepstral coefficients (MFCCs): MFCCs, which are obtained by applying cosine transform to a sub-band Mel-frequency spectrum within a short time, have been widely

used in bird classification [Lee et al., 2006], speech/speaker recognition [Han et al., 2006], and frog identification [Bedoya et al., 2014]. In this chapter, MFCCs are calculated based on the method of [Lee et al., 2006].

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

$$E_j = \sum_{k=0}^{N/2-1} \phi_j(k) A_k, 0 \leq j \leq J-1 \tag{3.14}$$

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

$$A_k = |X[k]|^2, 0 \leq k \leq N/2 \tag{3.15}$$

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

$$C_m^j = \sum_{j=0}^{J-1} \cos(m\frac{\pi}{J}(j+0.5)) log_{10}(E_j), 0 \leq m \leq L-1 \tag{3.16}$$

where L is the number of MFCCs.

In this chapter, 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 from each frame, the averaged MFCCs over all frames within one syllable are calculated as

$$f_m = \frac{\sum_{i=1}^{K}(C_m^l)}{K}, 0 \leq m \leq L-1 \tag{3.17}$$

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 concatenated with the other 11 feature vectors to form the hybrid temporal, perceptual and cepstral (TemPerCep) features.

After the formulation of feature vectors, normalisation is conducted as

$$v_i = \frac{v_i - \mu_i}{\sigma_i} \tag{3.18}$$

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 hybrid procedure is performed as

$$F_H = w_1 F_1 \oplus w_2 F_2 \oplus w_3 F_3 \tag{3.19}$$

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

3.2.5 Classifier description

In this chapter we report the classification results for five classifiers: 1) LDA, 2) kNN, 3) SVM, 4) RF, 5) ANN. Five feature sets, LPCs, MFCCs, combined temporal feature and MFCCs (*TemCep*), combined temporal and perceptual features (*TemPer*), combined temporal, perceptual features, and MFCCs (*TemCepPer*), are fed into each classifier respectively to test their classification performance.

Linear discriminant analysis

After transforming the feature vector into low-dimensional space, the classification accuracy can be improved for 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

$$J_F(A) = tr((A^T S_w A)^{-1}(A^T S_B A)) \tag{3.20}$$

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

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (F_i^j - \mu_j)(F_i^j - \mu_j)^T \tag{3.21}$$

$$S_B = \sum_{j=1}^{C} (\mu_i - \mu)(\mu_i - \mu)^T \tag{3.22}$$

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 $S_W^{-1}S_B$.

$$A_{opt} = argmax\frac{tr(A^T S_B A)}{A^T S_W A} \tag{3.23}$$

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.

K-nearest neighbour

For the kNN classifier, an object is classified to the majority class of its k nearest neighbours [Huang et al., 2009]. Specifically, in the training phase, frog feature vectors are stored with species labels. For the test phase, the simplest classification combination method is the voting method. The k closet vectors are selected for voting, then the classification for the input feature vector $f_{i,c}$ is assigned with the majority class.

The second classification combination method is to calculate the average distance between an input frog feature vector and k closest vectors. For example, the Euclidean distance between an input feature vector $f_{i,c}$ and one stored feature vector $f_{j,c}$ is calculated as

$$d(i,j) = \sqrt{\sum_{c=1}^{n}(f_{i,c} - f_{j,c})^2} \tag{3.24}$$

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

$$\frac{1}{k_1}\sum_{j\in s_1} d(i, j(s_1)) < \frac{1}{k_2}\sum_{j\in s_2} d(i, j(s_2)) \tag{3.25}$$

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.

The third classification combination method is to calculate the sum of similarity of k closest feature vectors. For a binary classification task with two classes: k_1 and k_2. If

$$\sum_{j \in s_1} d(i, j(s_1)) < \sum_{j \in s_2} d(i, j(s_2)) \tag{3.26}$$

Then the input feature vector i will be classified as belonging to class s_2. Following prior work ([Han et al., 2011, Xie et al., 2015b]), the distance function used for kNN is the Euclidean function, and k is set at 1.

Support vector machines

Due to the high accuracy and superior generalisation properties, SVM has been widely used for classifying animal sounds [Huang et al., 2009] [Acevedo et al., 2009]. In this chapter, 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 instances in the training data, F_l^n is the feature vector obtained from the $l\text{-}th$ 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 [Cortes and Vapnik, 1995] is defined by the training data as

$$f(v) = sgn(\sum_{sv} \alpha_l^n L_l^n K(v, v_l^n) + b_l^n) \tag{3.27}$$

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 [Hsu et al., 2003].

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 [Bao and Cui, 2005]. The prediction is defined as

$$Pred = \frac{1}{K}\sum_{n=1}^{K} T_i \tag{3.28}$$

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.

Artificial neural network

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

$$D(F) = argmaxp_i(F), i = 1, ..., N \tag{3.29}$$

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

$$p_i(F) = \sum_{j=1}^{m_i} \beta_{ij}\phi_{ij}(F) \tag{3.30}$$

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

$$\sum_{j=1}^{m_i} \beta_{ij} = 1 \tag{3.31}$$

$$\phi_{ij}(F) = \frac{1}{(2\pi)^{(}d/2)\sigma^d} exp[-\frac{(F-\mu_{ij})^T(F-\mu_{ij})}{2\sigma^2}] \tag{3.32}$$

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 chapter, one ANN classifier named MLP is used to classify frog calls.

3.3 Experiment results

In this experiment, performance statistics is evaluated using 5-fold cross-validation for testing the robustness of our proposed feature set. 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 found in Chapter 2.6.1.

3.3.1 Effects of different feature sets

Figure 3.5 illustrates the classification accuracy with different feature sets: LPCs, 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 LPCs and perceptual features. It is evident that our proposed combined feature set (*TemPerCep*) shows outstanding performance of all proposed feature sets of all the machine learning techniques. The reason for the high classification accuracy is that frog calls are of short duration and cover a small spectral band. Our proposed combined feature set, *TemPerCep*, can better characterise the content of frog calls. Although the classification performance of *TemPerCep* is not significantly higher than other feature sets, the difference does show that our proposed feature set is suitable and effective for the classification of frog calls.

3.3.2 Effects of different machine learning techniques

Figure 3.6 shows the frog call classification performance of $TemPerCep$ with different machine learning techniques. The high classification results in term of the accuracy, sensitivity and specificity measure of different classifiers indicates good classification performance. It can be observed that RF achieves the best classification performance, while the classification performance of LDA is the lowest. Meanwhile, the classification performance of SVM and

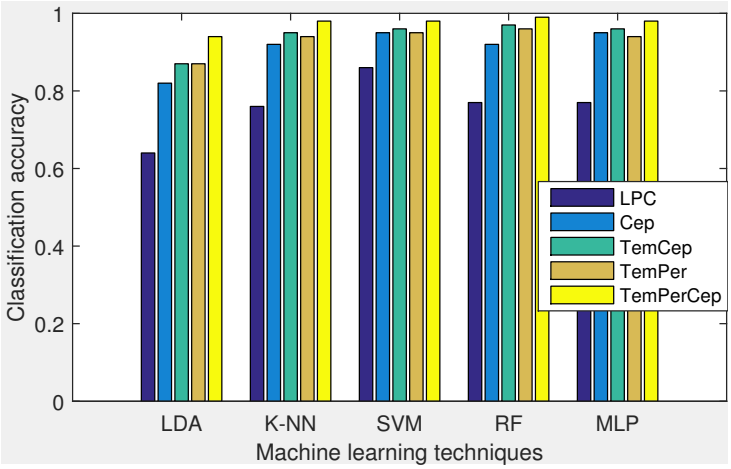


Figure 3.5: Classification results with different feature sets using the window size of 64 samples

MLP is very good, which might be that the features and classifiers are quite suitable. It can be seen from Figure 3.6 that frog call classification with different machine learning techniques can achieve good performance with our combined feature set, because the classification accuracy is very high. It can also be noted that RF can be highly recommended for classification of frog calls due to the highest classification accuracy.

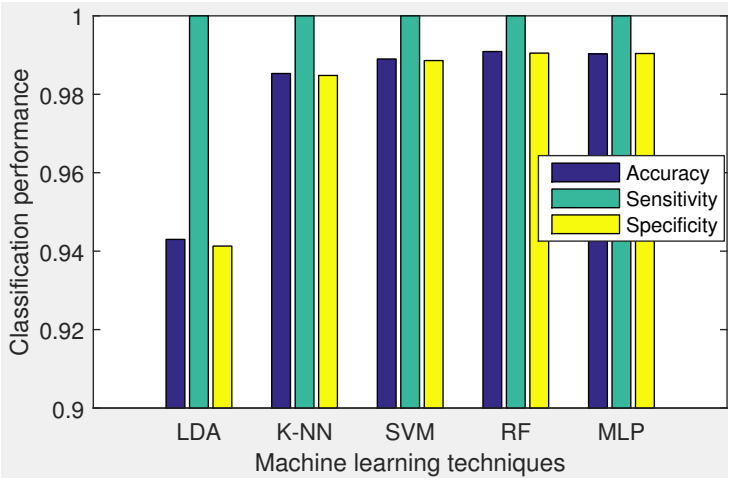


Figure 3.6: Results of different classifiers

3.3.3 Effects of different window size for MFCCs and perceptual features

As we know, the window size has an effect on the MFCCs and perceptual features. Therefore, a different window size will lead to different classification performance (Figures 3.7 and 3.8). The window size used for the test is 32, 64, 128, 256, 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 obtains the best classification performance. It also can be observed that SVM and RF achieve the best classification performance. Moreover, different window sizes of MFCCs have a larger variation than *TemPer* features, which might be because temporal features have a high weight in *TemPer* for the classification task.

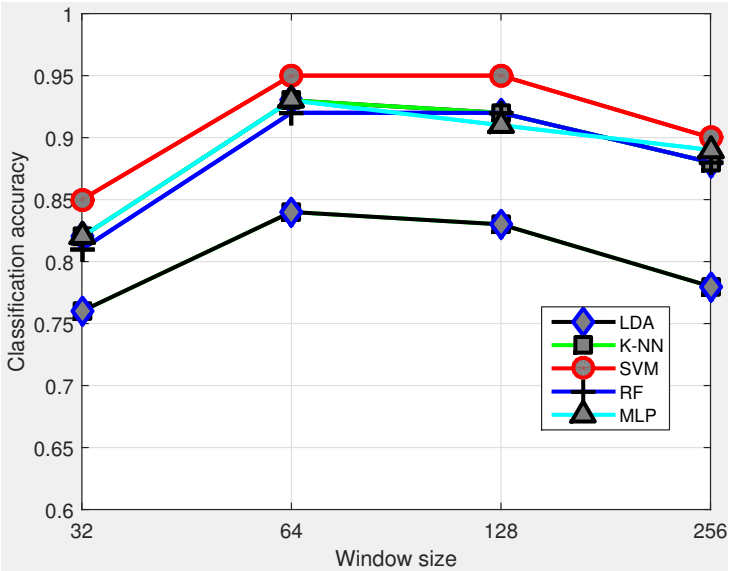


Figure 3.7: Classification results of MFCCs with different window sizes

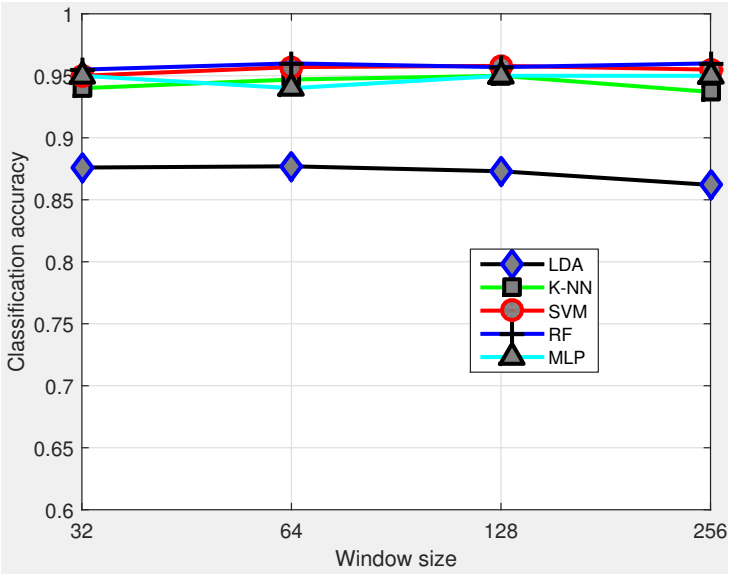


Figure 3.8: Classification results of TemPer with different window sizes

3.3.4 Effects of noise

To further evaluate the robustness of our proposed feature set, white noise with different SNRs of 40 dB, 30 dB, 20 dB, 10 dB, 0dB, and -10 dB is added to the frog calls. Because this chapter focuses on the evaluation of features rather than the segmentation method, the artificial noise is added after syllable segmentation. Since SVM has shown good performance for frog call classification in 3.3.1, we only use SVM to test the effects of different levels of artificial noise. The classification results of different levels of noise contamination are shown in Figure 3.9. It is found from Figure 3.9, that MFCCs (Cep) are very sensitive to background noise, compared to other feature sets. Comparing *TemCep* with *TemPer*, it can be observed that perceptual features have a better anti-noise ability than the cepstral feature. It is also found that LPCs have a good anti-noise ability when SNR is larger than 10 dB, but the classification accuracy quickly decreases when SNR is smaller than 10 dB.

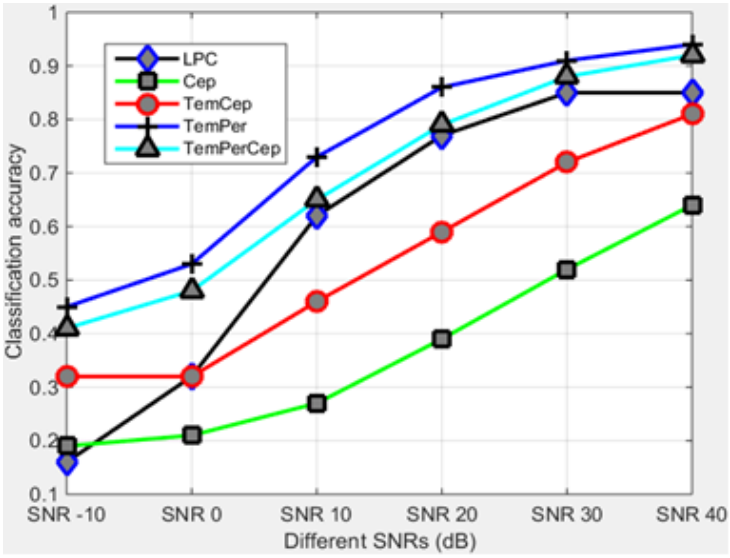


Figure 3.9: Sensitivity of different feature sets for different levels of noise contamination

3.4 Discussion

Table 3.2 shows the classification performance of previous methods. Since previous studies often used different datasets to perform the classification task, this research implements all those features and apples them to the used dataset with the same classifier (SVM). Compared to those previous methods, this proposed combined feature set significantly outperforms other methods. Therefore, it can be concluded that the results of this research stand above the current

classification performance. From Table 3.2, we can also observe that MFCCs are the most popular feature that has been used for frog call classification. Among all used machine learning techniques, SVM shows the superior performance and is widely used for the classification task. It can be found that the classification accuracy of *TemPerCep* does not show significant improvement when compared to MFCCs. However, combining temporal and perceptual features with cepstral features greatly improves the anti-noise ability of MFCCs.

Table 3.2: Comparison with previous used feature sets

Ref.	Feature	Accuracy (%)
[Juan Mayor, 2009, Yuan and Ramli, 2013]	LPCs	93.5
[Bedoya et al., 2014, Jaafar and Ramli, 2013, Lee et al., 2006, Xie et al., 2015b]	MFCCs	94.9
[Han et al., 2011]	Spectral centroid, Shannon entropy, Rényi entropy	75.6
[Xie et al., 2015b]	Syllable duration, dominant frequency, oscillation rate, frequency modulation, energy modulation	92.3
[Huang et al., 2014]	Spectral centroid, signal bandwidth, spectral roll-off, threshold-crossing rate, spectral flatness, and average energy	95.8
Our feature set	*TemPerCep*	99.1

3.5 Summary

In this chapter, a novel combined feature set was proposed to classify frog calls in trophy recordings with five machine learning algorithms. After segmenting continuous recordings into individual syllables, a variety of acoustic features are extracted from each syllable. Then, different features are combined to construct different feature sets. Finally, five machine learning techniques are used to classify frog calls in trophy recordings with different feature sets. Classification accuracy for 24 frog species in trophy recordings is 99.1%, which is much higher than other feature sets. The results demonstrate that a combination of temporal, spectral and cepstral features outperforms the state-of-the-art features used for frog call classification in trophy recordings. Compared to temporal and spectral features, cepstral features achieve a higher classification accuracy when used individually. However, they are sensitive to the background noise. Therefore, we aim to develop a novel cepstral feature with a good anti-noise ability in the subsequent analysis.

Chapter 4

Adaptive frequency scaled wavelet packet decomposition for frog call classification

---

Research problem

Following the summary of Chapter 3, cepstral features have been widely used for classifying frog species in trophy recordings with a high classification accuracy. However, they are very sensitive to background noise.

Research sub-question

How to develop robust cepstral features to classify frog species in both trophy and field recordings?

4.1 Overview

This chapter presents a novel cepstral feature based on adaptive frequency scaled wavelet packet decomposition (WPD), whose goal is to develop a novel feature with a good anti-noise ability. Since both trophy and field recordings are studied in this chapter, developing features with a good anti-noise ability is important for dealing with field recordings. Different from most previous studies that extracted features via Fourier transform, WPD is employed for feature extraction. The classification performance is evaluated with two different datasets from Queensland, Australia (high SNR: 18 frog species from trophy recordings and low SNR: field recordings of eight frog species from James Cook University recordings). Although trophy recordings are used in this chapter, each recording is assumed to have only one frog species and the classification framework is regarded as a SISL classification problem.

4.2 Methods

The architecture of the proposed classification method contains five modules: syllable segmentation, syllable feature extraction, adaptive frequency scale generation, WPD feature extraction and classification (see Figure 4.1). Each module is described in the following sections.

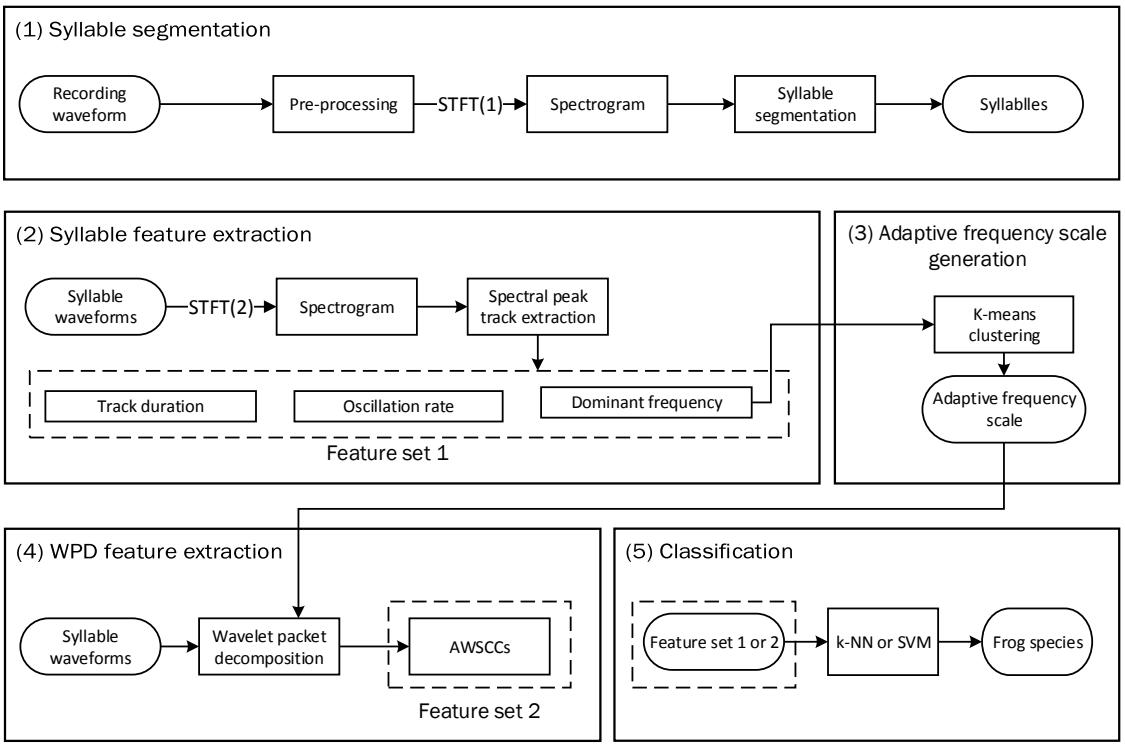


Figure 4.1: Block diagram of the frog call classification system. The line of dashes indicates the extracted feature set. AWSCCs is the abbreviation of *adaptive wavelet packet decomposition sub-band cepstral coefficients*. STFT is short-time Fourier transform. For STFT(1), the window function, size 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 Figure 4.6

4.2.1 Sound recording and pre-processing

Two datasets obtained from a trophy recording [Stewart, 1999] and James Cook University (JCU) were selected for this chapter. 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 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 described in

Figure 4.1. To reduce the subsequent computational burden, all the recordings selected from the CD were re-sampled at 16 kHz per second, mixed down to mono, and saved in WAV format.

The JCU recordings were obtained from Kiyomi dam (S $19^\circ\ 22^{'}\ 16.0^{''}$, E $146^\circ\ 27^{'}\ 31.3^{''}$) BG Creek dam (S $19^\circ\ 27^{'}\ 1.23^{''}$, E $146^\circ\ 24^{'}\ 5.65^{''}$) and Stony Creek dam (S $19^\circ\ 24^{'}\ 07.0^{''}$, E $146^\circ\ 25\ 51.3$) in Townsville, using acoustic sensors. The recordings were stored on 16 GB SD cards in 64 kbps MP3 mono format and have a low SNR compared to the trophy recording. The sample rate is 16.00 kHz. All the JCU recordings started around sunset, finished around sunrise every day and have 12 hour duration. In this chapter, JCU recordings are field recordings.

4.2.2 Spectrogram analysis for validation dataset

In this chapter, three syllables for each frog species are set aside and used as a *reference data set*. For the trophy 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 4.1. The reference data set is excluded from the data used in the testing stage.

Table 4.1: Parameters of 18 frog species averaged of three randomly selected syllable samples in the trophy recording. These selected samples make the *reference data set*

No.	Scientific name	Abbreviation	Syllable duration (millisecond)	Peak frequency (Hz)	Oscillation rate (cycle/second)
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

For the JCU recordings[2], the corresponding parameters are described in Table 4.2. Compared to the trophy recordings from the CD, peak frequency shows a smaller variation than syllable duration and oscillation rate.

Table 4.2: Parameters of eight 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 durations of *Rhinella marina* (Commom name: Canetoad) are very different from each other, we manually set the duration of Canetoad using the maximum duration of other frog species, which is 500 milliseconds

No.	Scientific name	Abbreviation	Syllable duration (millisecond)	Peak frequency (Hz)	Oscillation rate (cycles/second)
1	Rhinella marina	RMA	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

4.2.3 Syllable segmentation

The syllable segmentation process is described in Chapter 3.2.2. To further improve the segmentation result, the averaged energy of which 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 Figure 4.2.

For the JCU recordings, bandpass filtering is applied to each recording before using the H$\ddot{a}$rm$\ddot{a}$'s method [Härmä, 2003]. A bandpass filter is first used to filter specific frog species, because different frog species tend to call simultaneously. The filtering is

$$S^{'}(t,f) = \begin{cases} S(t,f) & F_{lower} \leq f \leq F_{upper} \\ 0 \end{cases}$$

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

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

frequency and calculated as

$$F_{upper} = F_{peak} + \beta$$
$$F_{lower} = F_{peak} - \beta \tag{4.1}$$

where F_{peak} is the peak frequency (Table 4.2), 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 result 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 the H*ä*rm*ä*'s method to detect individual syllables (Figure 4.3).

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

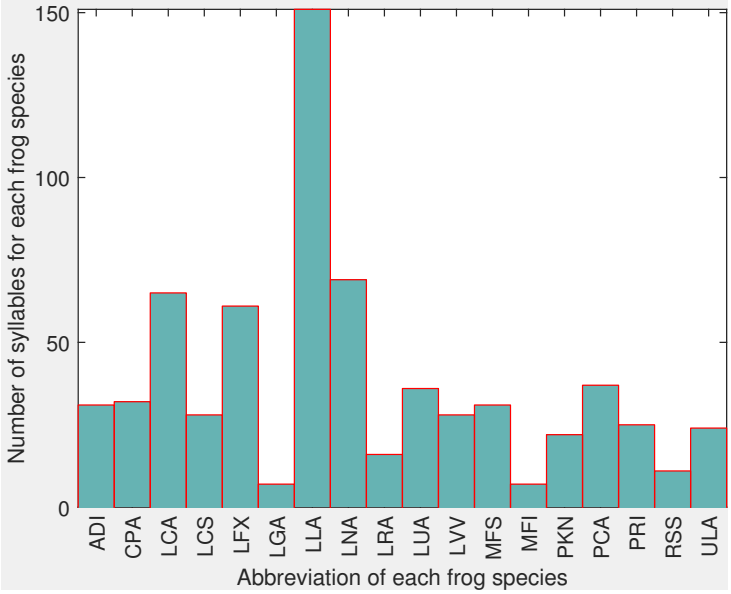
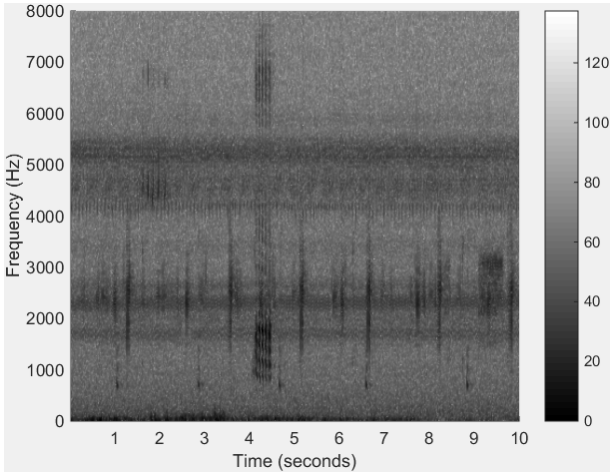


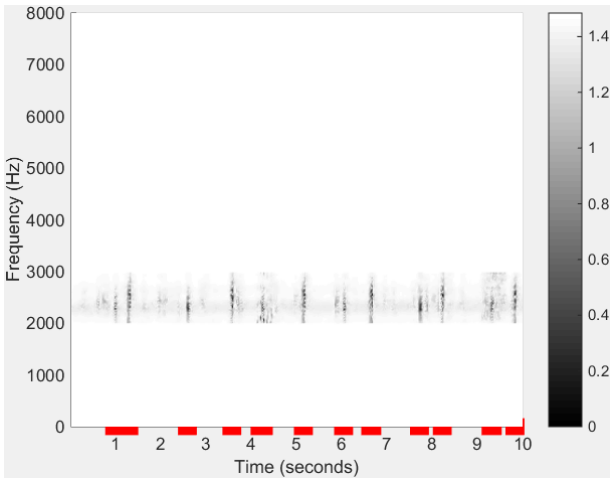
Figure 4.2: Distribution of syllable number for all frog species. The x-axis is the abbreviation of each frog species, and the corresponding scientific name can be found in Table 4.1

4.2.4 Spectral peak track extraction

Spectral peak tracks (SPT) (also called frequency tracks) have been explored for studying birds [Heller and Pinezich, 2008, Jancovic and Kokuer, 2015] and whales [Roch et al., 2011]. In this chapter, the spectral peak track is used to represent the trace of a frog advertisement call, because frogs that are genetically related share more similar advertisement calls than distantly



(a) Spectrogram.



(b) Segmentation results with marked red lines.

Figure 4.3: Segmentation results for *Uperolela mimula* using bandpass filtering, noise reduction and H*ä*rm*ä*'s method. The red line in (b) indicates the start and stop location of each segmented syllable

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 method is reported in [Xie et al., 2015b].

For the SPT extraction algorithm, seven parameters need to be set (Table 4.3). 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

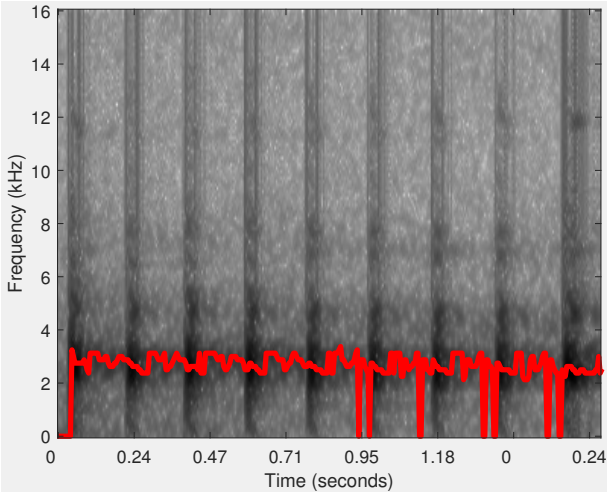
Table 4.3: Parameters used for spectral peak extraction

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$\sim$1)	Minimum density value

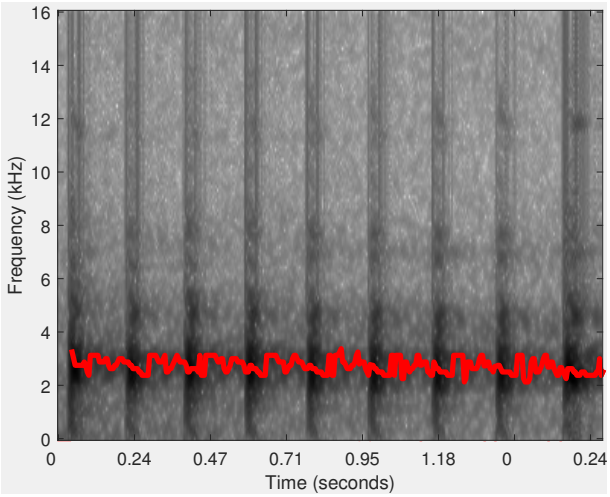
intervals are smaller than T_c and f_c respectively, one initial track (SPT_1) will be generated. After that, linear regression is applied to the generated track for calculating the position of the next predicted peak. Based on peaks $p_1(t_1, f_1)$ and $p_2(t_2, f_2)$ within the initial track (SPT_1), a and b in Equation (4.2) can be solved.

$$f = at + b \tag{4.2}$$

Based on a and b, the predicted peak $\hat{p_n}$ of the following frame t_n can be calculated. Next, the time and frequency domain intervals between predicted peak ($\hat{p_n}$) 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 Figure 4.4. 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 Figure 4.4.



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



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

Figure 4.4: Spectral peak track extraction results for *Neobatrachus sudelli*. By filling the gaps within the track, the dominant frequency can be more accurately calculated

4.2.5 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 ($t_s \leq t \leq t_e$). 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.

$$D_t = (t_e - t_s) * r_x \tag{4.3}$$

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

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

$$\bar{f} = \sum_{t=t_s}^{t_e} f_t/(t_e - t_s + 1) * r_y \tag{4.4}$$

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 is introduced and summarised 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 normalised, 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 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

$$O_r = \frac{P_f}{L_{dct}} * r_x * \gamma \tag{4.5}$$

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 set as 0.2 second in this experiment.

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

4.2.7 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 [Biswas et al., 2014, Litvin and Cohen, 2011, 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 chapter, an adaptive frequency scale for WPD for frog calls is proposed, 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 Figure 4.5.

4.2.8 Feature extraction based on adaptive frequency scaled WPD

In previous studies [Bedoya et al., 2014, Xie et al., 2015b], MFCCs have been used for studying bioacoustic data, and it is used as the baseline for feature comparison in this chapter. 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 chapter, 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 the Mel-scale for MWSCCs. Meanwhile, after performing DCT, temporal feature integration is used for calculating the

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

statistics of feature vectors which generates different statistical types of AWSCCs. (see in Figure 4.6).

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 $y(n)$.

$$x(n) = w(n)y(n) \tag{4.6}$$

where $w(L)$ is the Hamming window function and defined as $w(n) = 0.54 - 0.46cos(\frac{2n\pi}{L-1})$, 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 4.2.7.

$$WP(i, j) = \sum_{i=1}^{M} x(n)\psi_{(a,b)}(n) \tag{4.7}$$

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'

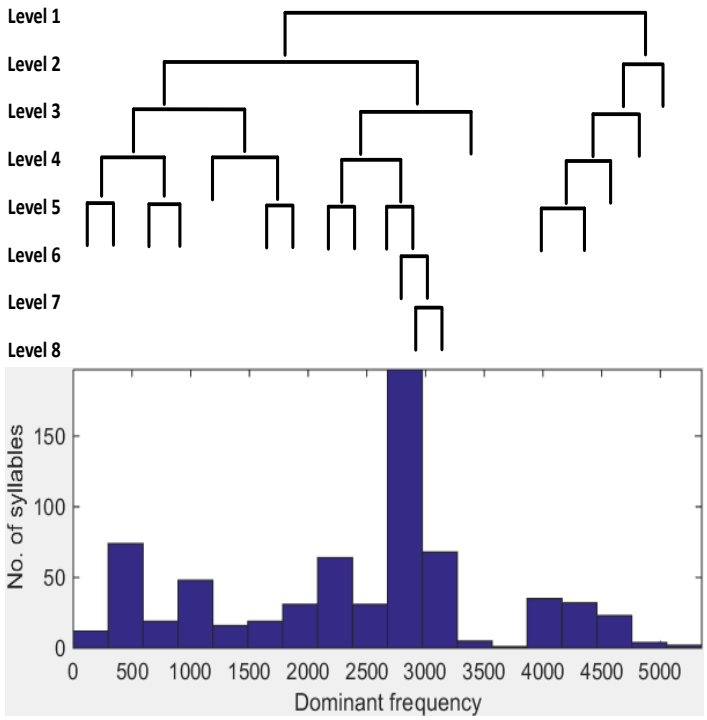


Figure 4.5: Adaptive wavelet packet tree for classifying twenty frog species. The upper image is the wavelet packet tree; the lower image is the histogram of dominant frequency for twenty frog species

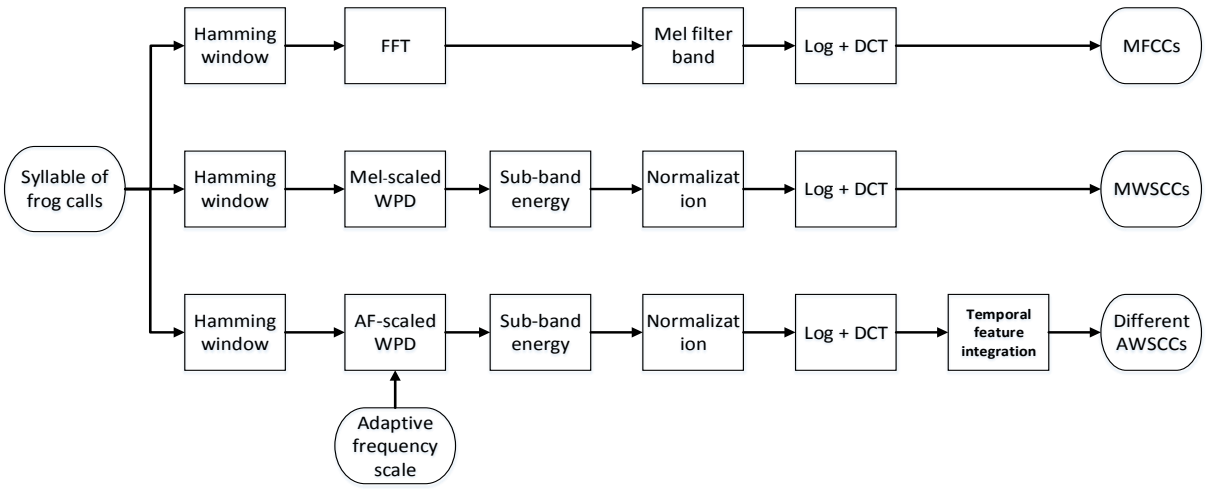


Figure 4.6: Description of three feature extraction methods including MFCCs, MWSCCs, and different statistical types of AWSCCs

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.

$$WP_i = \sum_{j=1}^{M_i} [WP(i,j)]^2 \tag{4.8}$$

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). Normalise the energy of each sub-band.

$$SE_i = \frac{WP_i}{M_i} \tag{4.9}$$

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

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

$$AWSCCs(d) = \sum_{i=1}^{T} logSE_i cos(\frac{d(i-0.5)}{T}\pi) \tag{4.10}$$

where $d = 1, 2, ..., d^{'}$, $1 \leq d^{'} \leq T$, 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 chapter, 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.

4.2.9 Classification

In this chapter, kNN and SVM classification algorithms 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. The descriptions of those two classifiers can be found in Chapter 3.2.5.

4.3 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 SNRs are described.

4.3.1 Parameter tuning

There are five modules that require parameter tuning: syllable segmentation, spectral peak track, feature extraction, and classification (Figure 5.1).

For syllable segmentation, the window size and overlap are 512 samples and 25%, but the intensity thresholds are 10 dB and 5 dB for the trophy recordings and the JCU recordings, respectively.

In the spectral peak track determination, there are seven parameters (see in Table 4.3). The parameter settings are shown in Table 4.4.

Table 4.4: Parameter setting for calculating spectral peak track

Parameter	Trophy 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$\sim$1)	0.8	0.6

With a random parameter setting start, an iterative loop is performed for a fixed range of each parameter as seen in Table 4.1 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 [Han et al., 2011, Huang et al., 2009, Xie et al., 2015b], the distance function used for kNN 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].

4.3.2 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 fed to two classifiers, which are the kNN and SVM classifiers. Due to the non-uniform distribution of the number of syllables for different frog species in the trophy recordings, a weighted classification accuracy is define as

$$weighted\,Acc = \sum_{i=1}^{N} Acc(i) * \frac{n_i}{N} \tag{4.11}$$

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.

4.3.3 Comparison between different feature sets

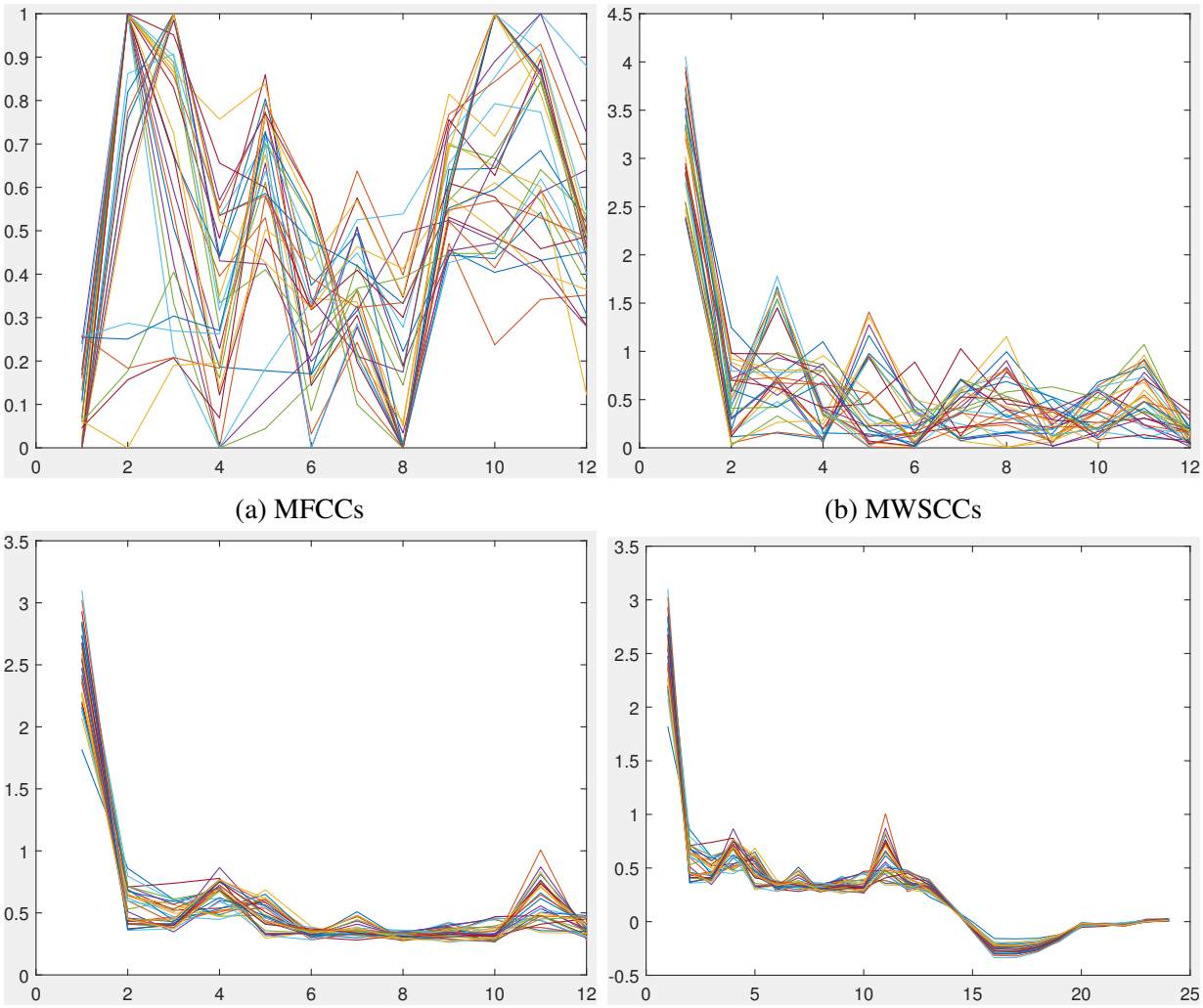
The classification accuracy comparison for 18 frog species using five feature sets and two classifiers is shown in Table 4.5.

Table 4.5: Weighted classification accuracy (mean and standard deviation) comparison for five feature sets with two classifiers

Feature set	**Classification accuracy (%)**	
	kNN	**SVM**
SFs	82.2 ± 11.2	84.2 ± 10.5
MFCCs	90.8 ± 8.6	92.8 ± 11.0
MWSCCs	95.0 ± 7.7	97.6 ± 5.7
Averaged AWSCCs	98.8 ± 4.2	99.0 ± 3.6
Delta-AWSCCs	99.2 ± 2.1	$\mathbf{99.6 \pm 1.8}$

In this experiment, the best classification accuracy is 99.6%, which is achieved by the delta-AWSCCs with the SVM classifier. Compared to the average AWSCCs, the delta-AWSCCs

achieved slightly better performance. 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 that of averaged AWSCCs and delta-AWSCCs with 96.3%. The improvement shows that the proposed adaptive frequency scale can capture more information about frog calls than the Mel-scale (Figure 4.7).



(a) MFCCs

(b) MWSCCs

(c) Averaged AWSCCs

(d) Delta-AWSCCs

Figure 4.7: The feature vectors for 31 syllables of the single species, *Assa darlingtoni*. The x-axis is the 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)

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 the proposed feature, the frog call classification accuracy of all frog species is shown in Table 4.6. However, only the features that use the SVM classifier are shown, because averaged accuracy of the kNN classifier (93.2%) is lower than the SVM classifier (94.64%).

Table 4.6: 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 4.6 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 to 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%; this 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 4.8, the classification accuracy of SFs and MFCCs is lower than the other three features, at only 84.2% and 92.8%, respectively.

The statistical significance of the results is shown in Table 4.7. 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.

Table 4.7: Paired statistical analysis of the results in Table 4.6. For the classification accuracy of each frog species, the paired Student t-test was conducted [Tanton, 2005]

Pairs	**t-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)

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 4.1). The wavelet packet tree for classifying ten and fourteen frog species is shown in Figure 4.8, 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 Figure 4.9). The classification results are shown in Table 4.8. Since the classification accuracy with averaged AWSCCs is very high for classifying ten and fourteen frog species, the delta-AWSCCs is not included in this experiment. Table 4.8 shows that averaged AWSCCs can achieve the highest classification accuracy for classifying different numbers 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.

Table 4.8: Classification accuracy (%) for different number of frog species with four feature sets

Number of frog species	**SFs**	**MFCCs**	**MWSCCs**	**Averaged AWSCCs**
18 frog species	84.2 ± 10.5	92.8 ± 11.0	97.6 ± 5.7	99.0 ± 3.6
14 frog species	89.6 ± 9.7	94.4 ± 8.5	99.2 ± 2.6	100.0 ± 0.0
10 frog species	94.6 ± 8.7	95.8 ± 8.6	100.0 ± 0.0	100.0 ± 0.0

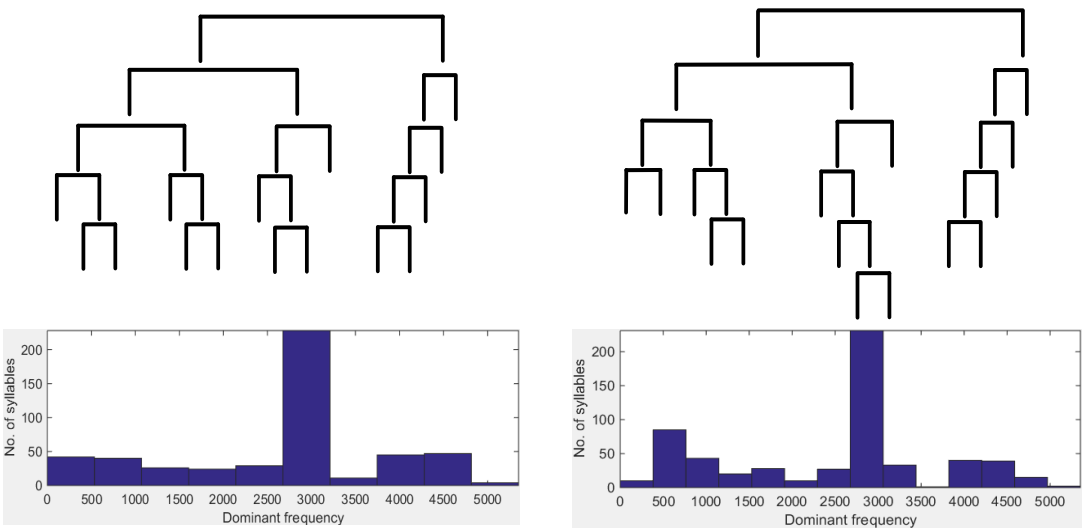


Figure 4.8: Wavelet packet tree based on adaptive frequency scale for classifying ten and fifteen frog species

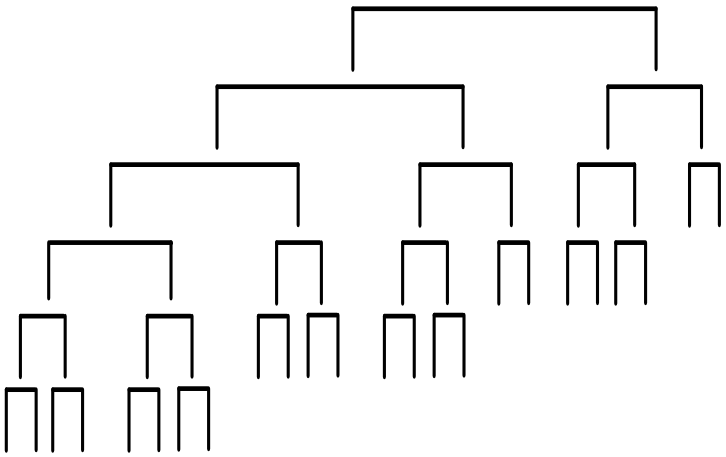


Figure 4.9: Mel-scaled wavelet packet tree for frog call classification

4.3.4 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 chapter 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 Figure 4.10. Compared to MFCCs and MWSCCs, SFs has 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 stronger anti-noise performance than other cepstral features (MFCCs and MWSCCs).

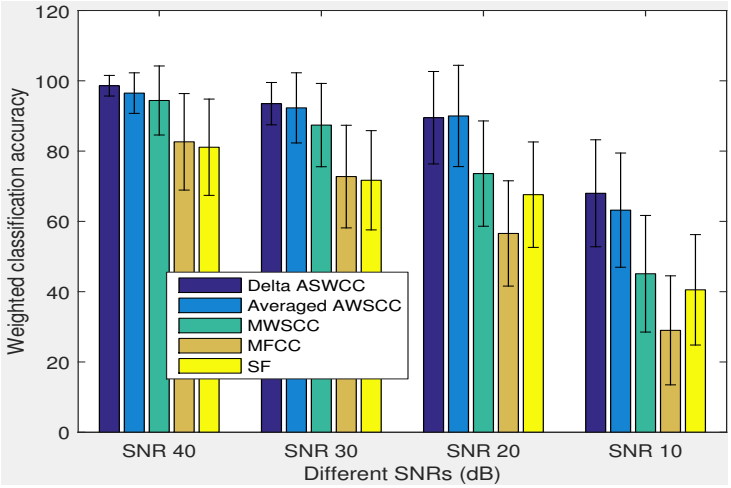


Figure 4.10: Sensitivity of five features for different levels of noise contamination

4.3.5 Feature evaluation using the real world recordings

Table 4.9 shows the classification accuracy comparison using our proposed feature to classify eight 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 to 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.

4.4 Summary

In this chapter, a novel feature extraction method for frog call classification is developed using the adaptive frequency scaled WPD. With segmented syllables, spectral peak track is first extracted from each syllable. Then, track duration, dominant frequency, and oscillation rate are

Table 4.9: Classification accuracy using the JCU recordings

Feature set	Classification accuracy (%)	
	kNN	**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	$\mathbf{97.4 \pm 5.4}$

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, AWSCCs, is calculated. Since the feature extraction method is developed based on the data itself, the wavelet packet tree varies according to the frog species to be classified. Compared to 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 characteristics of those frog species to be classified can be enhanced, while the background noise and calls from other animals will be suppressed. Therefore, the 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 efficient in classifying the frog calls when compared to Mel-scale (Figures 4.8 and 4.9).

Although both trophy and field recordings are used in this chapter, all the recordings are assumed to contain only one frog species per recording. In the next chapter, this limitation will be solved.

Chapter 5

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

---

Research problem

In Chapters 3 and 4, each individual recording is assumed to contain only one frog species. However, most field recordings collected by acoustic sensors have multiple frog species and a low SNR.

Research sub-question

How to classify multiple simultaneously vocalising frog species in field recordings?

5.1 Overview

This chapter proposes a method for the classification of multiple simultaneously vocalising frog species in field recordings. In Chapters 3 and 4, frog call classification is solved using a SISL classification framework, which cannot reflect the nature of automatically collected field recordings. Most field recordings have a low SNR and contain multiple simultaneously vocalising animals including frogs, birds, insects, and so on. This attribute makes MIML learning a natural fit for studying field recordings. Specifically, frog syllables in one audio clip (such as 10-second) are regarded as *multiple instance*, and frog species included in that audio clip denotes *multiple labels*. First, AED is used to segment frog syllables of each audio clip. Then, acoustic features are extracted from each segmented syllable. Lastly, three MIML

classifiers are used for classifying each 10-second recording.

5.2 Methods

Our MIML frog call classification framework contains four modules: signal processing, acoustic event detection, feature extraction, and classification (Figure 5.1). Detailed description of each module is listed in the following sections.

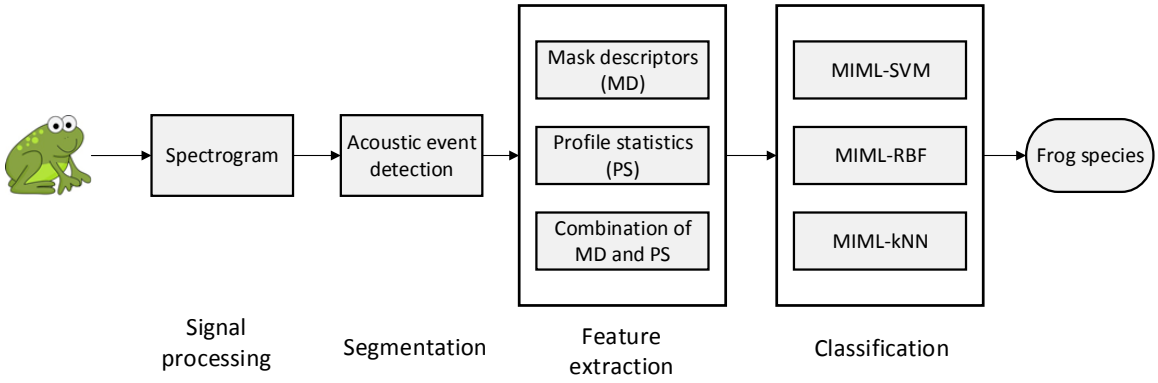


Figure 5.1: Flowchart of a frog call classification system using MIML learning

5.2.1 Materials

All recordings were obtained from three sites in Queensland, Australia: *Kiyomi dam*, *Stony creek dam* and *BG creek dam*. A battery-powered acoustic sensor (stored in a weather proof metal box) with an external microphone is used for the data collection. Collected recordings were stored on 16 GB SD cards in 64 kbps MP3 mono format. The recordings were collected from February, 2014 to April, 2014, because it is the breeding season in Queensland when male frogs make calls to attract females for the purpose of reproducing. Each recording starts around sunset, finishes around sunrise every day and has a duration of 12 hours. To evaluate our proposed MIML classification framework, a representative sample of 342 10-second recordings was prepared. The prepared recordings were manually labelled by an ecologist with eight frog species. Some recordings contains bird calls, insect calls, and numerous sounds, but all those sounds are regarded as the background noise. Each recording includes between one and six frog species. Following the prior work of [Briggs et al., 2012], we assume that recordings without frog vocalisations can be filtered out during the segmentation process. Acoustic parameters of

eight frog species averaged over three randomly selected syllables are shown in Table 4.2 of Chapter 4. Here, the acoustic parameters are used as the prior knowledge for event filtering.

5.2.2 Signal processing

Each recording was re-sampled at 16 kHz for generating a spectrogram using STFT. Specifically, each recording was divided into frames of 512 samples with 50% frame overlap. A fast Fourier transform was 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 calculated as

$$S_{tf} = 20 * log_{10}(A_{tf}) \tag{5.1}$$

where A denotes the amplitude value, $t = 0, ..., T-1$ and $f = 0, ..., F-1$ represent time and frequency index, T and F are 256 frequency bins and 625 frames, respectively.

5.2.3 Acoustic event detection for syllable segmentation

The aim of AED is to detect a specified acoustic event in audio data. In this chapter, we use AED for frog syllable segmentation. Since all recordings are collected from the field, there are many simultaneously vocalising frog species. Traditional methods for frog syllable segmentation are based on temporal information [Huang et al., 2009, Somervuo et al., 2004], which cannot address those environmenal recordings. Here, we modified the AED method developed by Towsey et al. [2012] for syllable segmentation. The detail of our AED method is described as follows.

Step 1: Wiener filtering

A 2-Dimensional Wiener filter is applied to the spectrogram image. A 5×5 pixel window is found to offer a satisfactory compromise between removal of background graininess and blurring of acoustic events. Wiener filtering aims to reduce the number of very small randomly distributed acoustic events appearing in the final output.

Step 2: Spectral subtraction

Wiener filtering can successfully remove the graininess, but some noises, such as wind, insect, motor engine that cover the whole recording, cannot be addressed. Here, spectral subtraction is used to deal with those noises (see Algorithm 2).

Algorithm 2: 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

Step 3: Adaptive thresholding

After noise reduction, the next step is to convert a noise reduced spectrogram $\hat{S_{tf}'}$ into the binary spectrogram S_{tf}^{b} for the event detection. Here, an adaptive thresholding method named *Otsu thresholding* [Otsu, 1975] is employed to find an optimal threshold.

$$\phi_b^2(k) = w_1(k)w_2(k)[\mu_1(k) - \mu_2(k)]^2 \tag{5.2}$$

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 normalised 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 j-th histogram bin. Then, the threshold, T_0, is calculated as

$$T_0 = (\phi_{b1}^2(k) + \phi_{b2}^2(k))/2 \tag{5.3}$$

Step 4: Events filtering using dominant frequency and event area

Since not all detected events correspond to frog vocalisations, those events that are not from the listed frog species in Table 4.2 of Chapter 5.2.1, dominant frequency (F_0) and area of the event (Ar) are used for filtering.

Step 5: Region growing

A region growing algorithm is to obtain the contour of each segmented acoustic event [Mallawaarachchi et al., 2008]. To get the accurate boundary of each acoustic event and improve the discrimination

of extracted features, a 2-dimensional region growing algorithm is used to obtain the accurate event shape for each segmented event. First, the point with the maximal intensity value within the event area is selected as the seed. Then, the neighbourhood pixels of the seed(s) above the threshold are located and assigned to the output image, and new added pixels are used as seeds for further processing. Finally, when all the pixels that satisfy the criteria are added to the output image, the recursive algorithm will stop and get the final results (Figure 5.3). Here, the threshold value is empirically set as 5 dB.

Algorithm 3: 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 2000 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 as 200 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.

5.2.4 Feature extraction

To compute features for each segment, the spectrogram and mask will be first cropped to contain just one segment. Figures 5.3(c) and 5.3(d) show a cropped image of the spectrogram and mask based on the highlighted segment. To describe the segment features, notation will be used as follows: Let $M_{t,f}$ be the cropped, binary mask for a segment, and let $\bar{S}_{t,f}$ be the cropped, original spectrogram. Note that t ranges from 1 to the duration of the segment in frames, T.

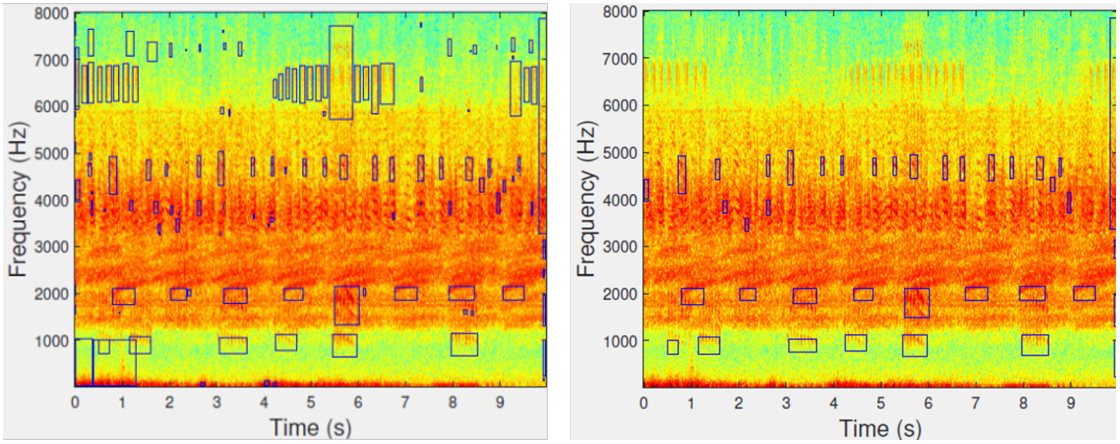


Figure 5.2: Acoustic event detection results before (Left) and after (Right) event filtering based on dominant frequency. Here, blue rectangle denotes the time and frequency boundary of each detected event

Two feature sets are calculated to describe each segment (syllable): mask descriptors and profile statistics [Briggs et al., 2012]. Here, we exclude *histogram of orientation (HOG)* from our feature set for its poor classification performance in previous studies [Briggs et al., 2012, Ruiz-Munoz et al., 2015].

Mask descriptors

Mask descriptors for a segment are based on only the mask, and describe the shape of the segment. The features are

(1) Min-frequency = min {f: $M_{t,f}$ = 1 }

(2) Max-frequency = max {f: $M_{t,f}$ = 1 }

(3) Bandwidth = max-frequency - min-frequency

(4) Duration = T

(5) Area = $\sum_{tf} M_{tf}$

(6) Perimeter = $\frac{1}{2}\times$ (# of pixels in $M_{t,f}$ such that at least one pixel in the surrounding 3 $\times$ 3 box is 1 and at least one pixel is 0)

(7) Non-compactness = $perimeter^2/area$

(8) Rectangularity = $area/(bandwidth \times duration)$

Profile statistics

Profile statistics are calculated based on statistical properties of the time and frequency profiles of each segment. To compute the time or frequency file, the columns or rows of the spectrogram are first summed. The time profile is $p_t(t) = \sum_f \hat{S}_{t,f}$ and the frequency profile is $p_f(f) = \sum_t \hat{S}_{t,f}$. The profiles are normalised to sum to 1, and are further interpreted as probability mass

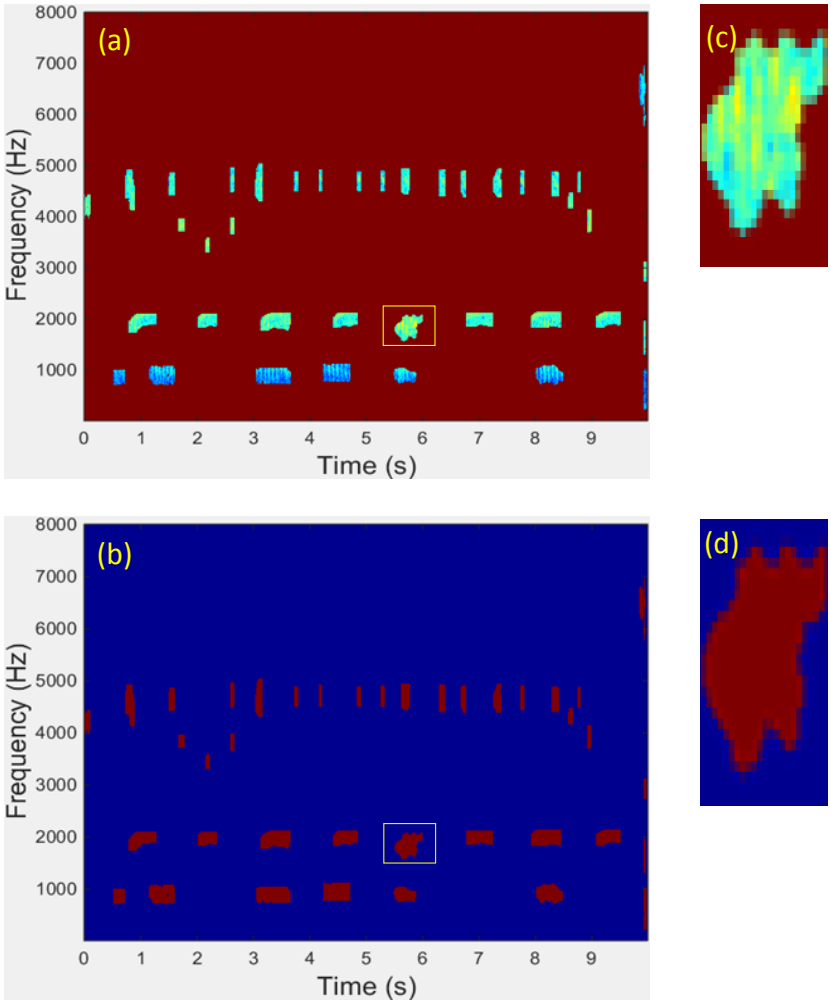


Figure 5.3: Acoustic event detection results after region growing. (a) binary segmentation results, (b) segmented frog syllables. (c) The masked and cropped spectrogram corresponding to the highlighted segment. (d) A cropped mask of the highlighted segment

functions. The normalised profile densities are $\hat{p}_t$ and $\hat{p}_f$. Two features measure the uniformity of the densities according the Gini index are

(1) Freq-gini = 1 - $\sum_f \hat{p}_f(f)^2$

(2) Time-gini = 1 - $\sum_t \hat{p}_t(t)^2$

Several more features are calculated by computing the k-th moments of the time and frequency profiles. Since different segments have different durations, all those features are calculated in a re-scaled coordinate system, where time is from 0 to 1 over the duration of the segment, and frequency is from 0 to 1.

(3) Freq-mean = μ_f = $\sum_{f=1}^{f_{max}} \hat{p}_f(f)(f/f_{max})$

(4) Freq-variance = $\sum_{f=1}^{f_{max}} \hat{p}_f(f)(\mu_f - f/f_{max})^2$

(5) Freq-skewness = $\sum_{f=1}^{f_{max}} \hat{p}_f(f)(\mu_f - f/f_{max})^3$

(6) Freq-kurtosis = $\sum_{f=1}^{f_{max}} \hat{p}_f(f)(\mu_f - f/f_{max})^4$

(7) Time-mean = $\mu_t = \sum_{t=1}^{t_{max}} \hat{p}_t(t)t/T)$

(8) Time-variance = $\mu_t = \sum_{t=1}^{t_{max}} \hat{p}_t(t)t/T)^2$

(9) Time-skewness = $\mu_t = \sum_{t=1}^{t_{max}} \hat{p}_t(t)t/T)^3$

(10) Time-kurtosis = $\mu_t = \sum_{t=1}^{t_{max}} \hat{p}_t(t)t/T)^4$

Also, the maxima of the time and frequency profiles are calculated

(11) Freq-max = $(argmax\hat{p}_f(f))/f_{max}$

(12) Time-max = $(\hat{p}_f(t))/T$

The mean and standard deviation of the spectrogram within the masked region are further calculated.

(13) Mask-mean = $\mu_{tf} = (1/area)\sum_{tf} \hat{S}_{tf}$

(14) Mask-stddev = $\sqrt{(1/area)\sum_{tf}(\mu_{tf} - \hat{S}_{tf})^2}$

Besides mask descriptors (MD) and profile statistics (PS), a third feature set is constructed with all features.

All of the features used to describe the segment are concatenated to form a single feature vector, but the values differ widely from each other. This property will affect the classification performance especially the distance-based classifier, where more weight will be placed on features with larger magnitudes. To prevent this bias, all those features are rescaled to [0,1] independently.

5.2.5 Multiple-instance multiple-label classifiers

After feature extraction, three MIML algorithms are evaluated for the classification of multiple simultaneous vocalising frog species: MIML-SVM, MIML-RBF, and MIML-kNN. These algorithms reduce the MIML problem to single-instance multiple-label problem by associating each bag (10-second recording) with a bag-level feature, which aggregates information from the instances in the bag [Briggs et al., 2012]. Each algorithm constructs different bag-level features, but all use some form of bag-level distance measure. Here, the maximal and average Hausdorff distances between two syllables are used by MIML-SVM and MIML-RBF, respectively. For MIML-kNN, the nearest neighbour is used to assign bag-level features.

5.3 Experiment results

5.3.1 Parameter tuning

There are three modules, the parameters of which 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 the constructed validation dataset. 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. Note that the dataset used for parameter setting is excluded from the testing dataset.

5.3.2 Classification

In this chapter, all the algorithms are programmed in Matlab 2014b. Each MIML algorithm is evaluated with five-fold cross-validation on the collection of 342 species-labelled recordings. Five evaluation metrics are used for comparing the performance with the combination of three feature sets and three MIML algorithms: Hamming loss, Rank loss, One error, coverage, and average precision [Madjarov et al., 2012a, Zhou et al., 2008]. The value range of all five evaluation rules is between 0 to 1. The definition of each evaluation rule is described as follows:
(1) Hamming loss is defined as the fraction of labels that are incorrectly predicted for an instance and the normalised Hamming loss which is normalised over instances is reported. This metric is defined as

$$hammingLoss = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{Q}|h(x_i)\Delta y_i)| \tag{5.4}$$

where Δ denotes the symmetric difference between two instances, N is the number of instances and Q is the total number of possible labels. y_i denotes the ground truth of instance x_i, and $h(x_i)$ denotes the predictions for the same instance.
(2) Ranking loss evaluates the average fraction of label pairs that are reversely ordered for the

particular instance given by

$$rankingLoss = \frac{1}{N}\sum_{i=1}^{N}\frac{|D_i|}{|y_i||\bar{y_i}|} \tag{5.5}$$

where $D_i = (\lambda_m, \lambda_n)|f(x_i, \lambda_m) \leq f(x_i, \lambda_n), (\lambda_m, \lambda_n) \in y_i \times \hat{y}_i$, while $\bar{y}$ denotes the complementary set of y in L, and $L = \lambda_1, \lambda_2, \lambda_3, ..., \lambda_Q$, λ represents the label.

(3) One error evaluates how many times the top-ranked label is not in the set of relevant labels of the instance. This evaluation metric is defined as

$$oneError = \frac{1}{N}\sum_{i=1}^{N}[[argmax_{\lambda \in y} f(x_i, \lambda)] \notin y_i] \tag{5.6}$$

(4) Coverage evaluates how far, on average, we need to go down the list of ranked labels in order to cover all the relevant labels of the example. The definition of this metric is shown as

$$coverage(f) = \frac{1}{N}\sum_{i=1}^{N} maxrank_f(x_i, \lambda) - 1 \tag{5.7}$$

where $rank_f(x_i, \lambda)$ maps the outputs of $f(x_i, \lambda)$ for any $\lambda \in L$ to $\lambda_1, \lambda_2, ..., \lambda_Q$, so that $f(x_i, \lambda_m) \leq f(x_i, \lambda_n)$ implies $rank_f(x_i, \lambda_m) \leq rank_f(x_i, \lambda_n)$

(5) Average precision is the average fraction of labels that are ranked higher than an actual label belonging to an instance.

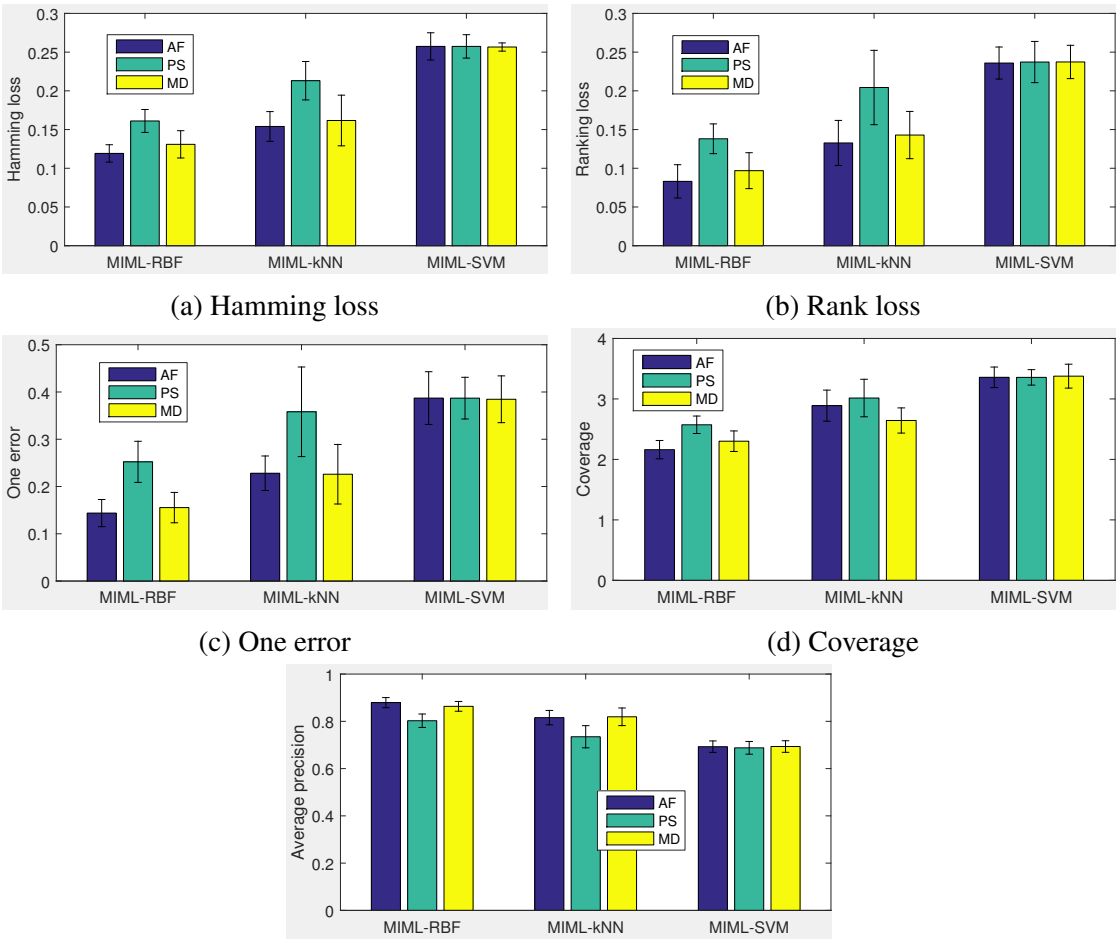
$$avgPrecision = \frac{1}{N}\sum_{i=1}^{N}\frac{h(x_i) \cap y_i}{|y_i|} \tag{5.8}$$

The values for hamming loss, rank loss, one-error, coverage, and average precision range from 0 to 1. For hamming loss, rank loss, one-error, and coverage, 0 denotes the perfect result, and 1 means the wrong prediction of all labels over every instance, whereas for average precision, the values have the completely opposite meanings.

5.3.3 Results

The classification results are shown in Figure 5.4. To obtain a base line for hamming loss, a non-informative classifier is always considered to predict the empty set. The baseline of

hamming loss is thus calculate as m/c, where c is the number of frog species to be classified, m is calculated as $(1/n)\sum_{i=1}^{n}|Y_i|$. Here, the value of baseline of hamming loss is 0.3220. It can be found that the hamming loss for MIML-RBF with AF is 2.70 times better than the non-informative classifier. With a rank loss of 0.0831, MIML-RBF with AF is 6.01 times better than the non-informative classifier. A one error of 0.1438 means that if we only predict the highest scoring species in each recording, it will truly be present 85.62% of the time. The $positive/negative$ is defined as $1 - hammingLoss$ and it is 88.08% for MIML-RBF with AF. Compared to MIML-kNN and MIML-SVM, MIML-RBF is found to achieve the best classification performance. For those three feature sets, the hamming loss for AF is always better than PS and MD.



(a) Hamming loss

(b) Rank loss

(c) One error

(d) Coverage

(e) Average precision

Figure 5.4: Evaluation metrics for MIML classifiers with three different feature sets

To give a concrete view of predictions, the results of five randomly selected recordings using MIML-RBF are shown in Table 5.1. From the table, we can see that recordings of No.4 are accurately predicted. Recordings of No.1, No.2, and No.3 are partially accurate. Recording

of No.5 is wrongly predicted.

Table 5.1: Example predictions with MIML-RBF using AF

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

In this chapter, all the features are directly calculate based on AED results. It is obviously that different AED results will lead to different classification results. A comparison of three AED methods is shown in Table 5.2, where features and classifier used are the same. The results show that our proposed AED method can achieve the best classification performance.

Table 5.2: Effects of AED for the MIML classification results. Here, ↓ indicates that smaller values imply higher accuracy, while ↑ has the completely opposite meanings

AED	Feature	Classifier	Hamming loss ↓	Coverage ↓	Average Precision ↑
Ours	AF	MIML-RBF	$\mathbf{0.1192 \pm 0.0112}$	$\mathbf{2.1614 \pm 0.1516}$	$\mathbf{0.8793 \pm 0.0213}$
Michael	AF	MIML-RBF	0.1275 ± 0.0090	2.2690 ± 0.1177	0.8529 ± 0.0217
Fodor	AF	MIML-RBF	0.1777 ± 0.0131	2.5471 ± 0.1339	0.8265 ± 0.0287

5.4 Discussion

Since most recordings in this chapter contain multiple simultaneously vocalising frog species, the traditional SISL classification framework is no longer suitable. A novel framework for the classification of multiple simultaneous vocalising frog species in field recordings is proposed, which is adopted from [Briggs et al., 2012], a study on birds. Different from their work, this research designs a new AED method for frog syllable segmentation rather than using a supervised learning algorithm. It is because that there are few annotated frog recordings. Since all the features in this study are calculated from the segmented syllables, the accuracy of the segmentation results directly affects the final classification performance. Compared to other two AED methods, extracting features based on our AED results can achieve better classification performance (Table 5.2).

We also investigate SISL classification results using a small dataset. Total 176 10-second recordings are annotated using eight frog species. On the average, each 10-second recording has 19.5 frog syllables. For SISL classification, MFCCs and AWSCCs are used as the features, respectively. The window size and overlap for calculating MFCCs and AWSCCs are 128 samples and 50%, which are selected according to Chapter 3. RF is used as the classifier for its best classification performance in Chapter 3. The MIML classification results are achieved using the combination of AF and MIML-RBF. The precision and recall for MIML and SISL classification are shown in Figure 5.5. Generally, MIML can achieve better performance than SISL for all frog species except for *RMA*. The high classification performance for *RMA* using SISL might be that frequency structure of *RMA* is much clearer than other frog species. Among all frog species, precision and recall are higher than 0.6 for MIML. Both precision and recall of *LNA* and *UMA* are 1, which are much higher than other frog species. The reasons for this high classification performance might be that the cepstral domain information of *LNA* and *UMA* can be better described by MFCCs. For SISL, *LUA* and *LFX* has the poorest classification performance. The reasons for poor classification accuracy of *LFX* might be insufficient training instances and inappropriate features for this frog species. Compared to MFCCs, our propose AWSCCs in Chapter 4.

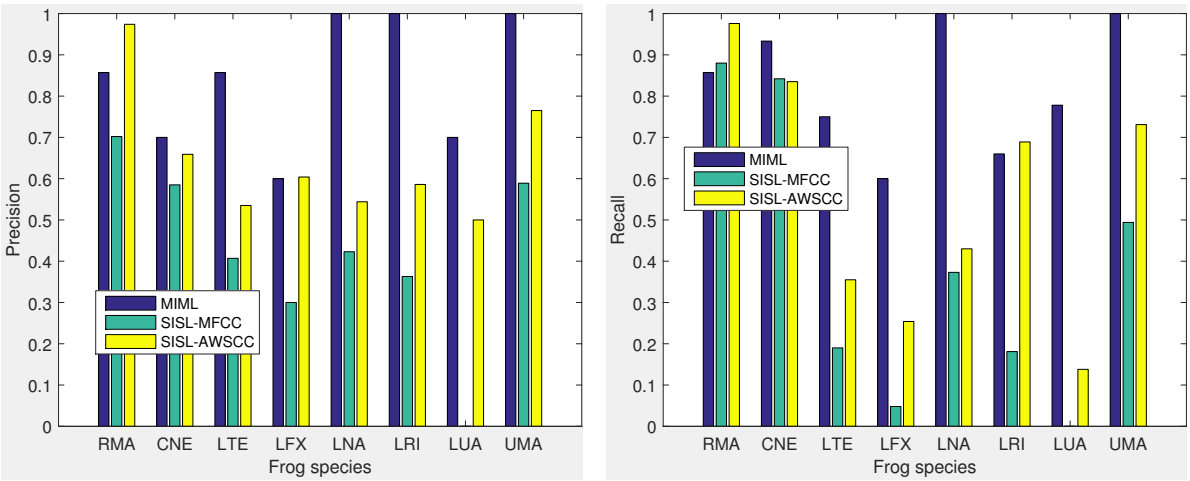


Figure 5.5: Comparisons of precisions and recalls between SISL and MIML classification for eight frog species

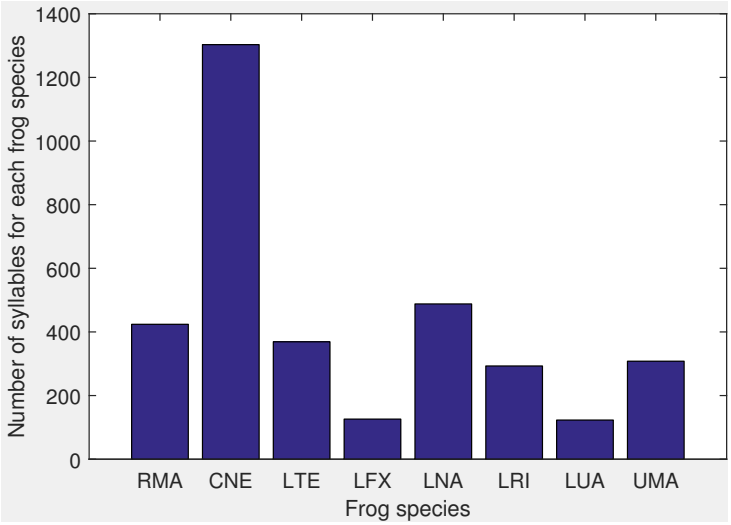


Figure 5.6: Distribution of syllable number for all frog species. The x-axis is the abbreviation of each frog species, and the corresponding scientific name can be found in Table 4.2

5.5 Summary

In this chapter, we propose a novel MIML classification framework for classifying multiple simultaneously vocalising frog species in field recordings. To the best of this author's knowledge, this is the first study that focuses on frog recordings using MIML learning. Since multiple frog species tend to call simultaneously, MIML learning is a natural fit for dealing with those recordings than SISL learning. For frog syllable segmentation, our propose AED method can achieve the best performance, which can be reflected by the classification results. Compared to SISL classification, MIML classification can achieve higher classification performance. Current classification results are found to be highly affected by the syllable segmentation results, and the use of AED cannot accurately segment all the syllables. One solution is to prepare an annotated dataset and apply supervised learning algorithms for the segmentation task. Another is to use a different classification framework, which does not need the segmentation process, and we examine this option in the next chapter.

Chapter 6

Frog call classification based on multi-label learning

Research problem

In Chapter 5, the classification performance is highly affected by the syllable segmentation results, which is realised by acoustic event detection (AED).

Research sub-question

How to reduce the effect of AED results and classify multiple simultaneously vocalising frog species in low SNR recordings?

6.1 Overview

This chapter describes the research conducted for classifying multiple simultaneously vocalising frog species. In Chapter 5, acoustic features are calculated based on acoustic event detection (AED) results, but the multiple-instance multiple-label (MIML) classification performance is highly affected by the accuracy of AED results. To reduce the bias introduced by AED, this chapter uses global feature sets for classifying multiple frog species in field recordings. To be specific, three features are calculated: linear prediction coefficients (LPCs), Mel-frequency cepstral coefficients (MFCCs), and adaptive-frequency scaled wavelet packet decomposition sub-band cepstral coefficients (AWSCCs). Here, each feature is extracted from the whole 10-second recording without syllable segmentation. Two cepstral feature sets are constructed by statistical analysis and spectral clustering. Since each 10-second recording is represented by a whole feature set and has multiple frog species, the classification process can be naturally

framed as a multiple-label (ML) learning framework.

6.2 Methods

6.2.1 Acquisition of frog call recordings

To evaluate the proposed algorithm, the same dataset with Chapter 5 is used. The description of this dataset can be found in Chapter 5.2.1. We first manually inspect spectrograms of ten randomly selected call examples for each frog species. Dominant frequency of each frog species as listed in Table 4.2 is used as prior information for subsequent analysis.

6.2.2 Feature extraction

Extracting discriminating features, which maximise between-group (inter-specie) dissimilarity and minimise within-group (intra-specie) dissimilarity, is very important for achieving high classification performance [Bedoya et al., 2014, Huang et al., 2009]. In this chapter, three global features are calculated to classify multiple simultaneously vocalising frog species in each 10-second recording: LPCs, MFCCs, and AWSCCs.

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

$$X(n) = \sum_{i}^{p} a_i x(n-i) \tag{6.1}$$

where p is the order of the polynomial a_i. In the proposed study, the value of p is set at 12 (12th-order polynomial), and 13 LPCs coefficients are calculated. For those frog vocalizations with different spectral envelopes, LPCs can obtain a high classification accuracy, and has been widely used in previous studies Jaafar and Ramli [2015], Jaafar et al. [2013a], Yuan and Ramli [2013].

MFCC: The description for calculating MFCCs can be found in Chapter 3.

AWSCCs: To calculate AWSCCs, constructing a suitable frequency scale for a 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 [Gingras and Fitch, 2013]. In Chapter 4, 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 chapter, the prior information for dominant frequency (F_0) (see Table 4.2) is directly used to construct the WP tree. Then, the steps for calculating AWSCCs are the same with Chapter 4.

6.2.3 Feature construction

Cepstral features are calculated of each frame, where each windowed signal contains n 12-dimensional feature vectors. Then, two methods are used to compute a reduced set of features.

The first method is to compute six statistical values for representing n 12-dimensional MFCCs [Dufour et al., 2013a]. Let MFCCs of each windowed signal be $V_i i = 1, ..., 12$, d and D are used to represent the velocity and acceleration of V.

then the six statistical values are calculated as follows:

$$f_1 = \frac{\sum_{i=1}^{n} V_i}{n} \tag{6.2}$$

$$f_2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(V_i - f_1)^2} \tag{6.3}$$

$$f_3 = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(d - d_i^2} \tag{6.4}$$

$$f_4 = \sqrt{\frac{1}{n-3}\sum_{i=1}^{n}(D - D_i)^2} \tag{6.5}$$

$$f_5 = \frac{\sum_{i=1}^{n-1} |d_i|}{n-1} \tag{6.6}$$

$$f_6 = \frac{\sum_{i=1}^{n-2} D_i}{n-2} \tag{6.7}$$

Finally, each windowed signal is represented as the concatenation of the six above features for the 12 cepstral coefficients, where the dimension of our feature set is 72.

The second method first clusters the 12 cepstral coefficients of all the windowed signal. Here, k-means clustering is used to reduce the preliminary dimension of all 12 cepstral coefficients. Then, dynamic time warping is used to calculate the distance between each clustered cepstral coefficients. Finally, spectral clustering is applied for further dimension reduction and getting the final feature vector. In this chapter, the value of k for k-means clustering is 50, and the number of clusters for spectral clustering is experimentally set at 2. Finally, the dimension of our constructed feature set is 24.

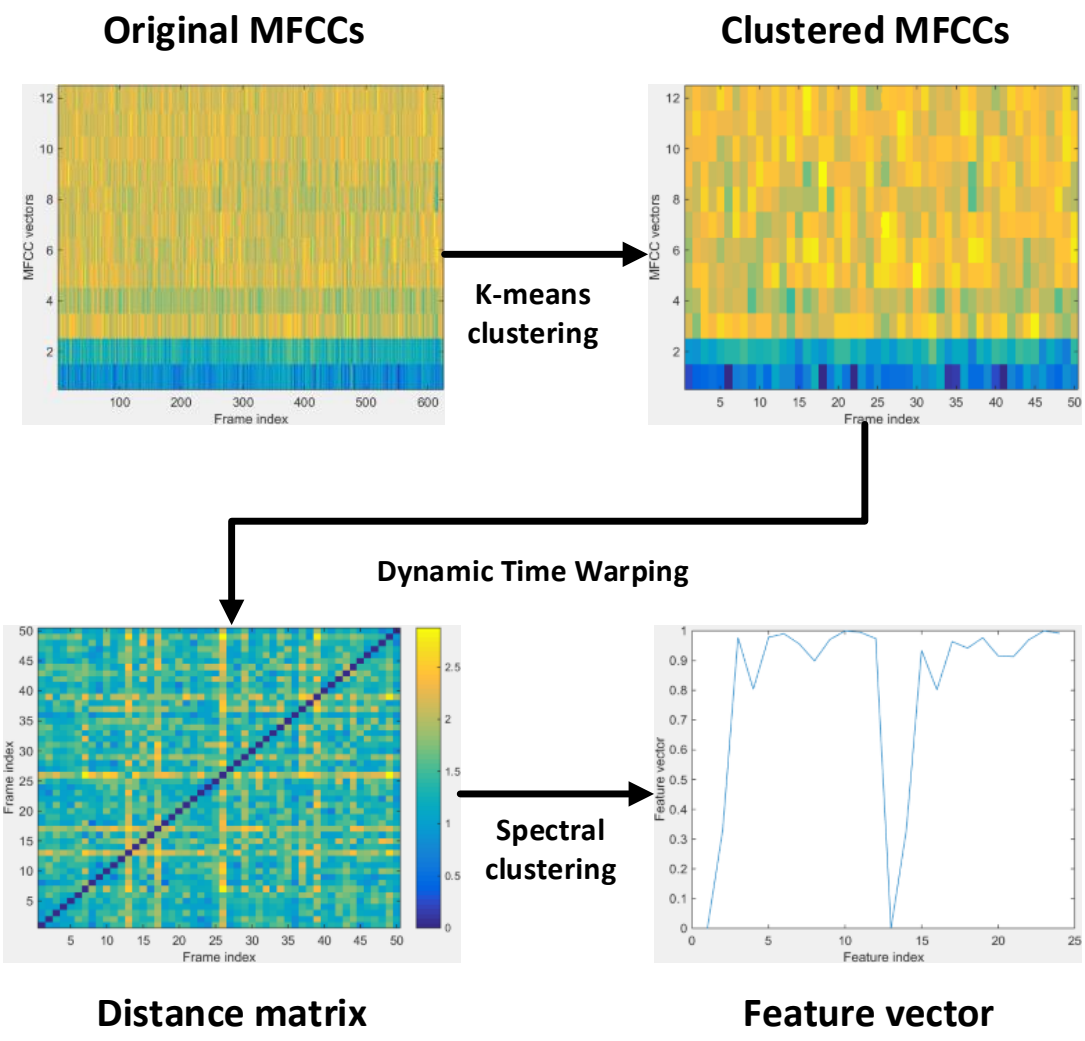


Figure 6.1: Procedure for extracting cepstral based feature vectors

6.2.4 Multi-label classification

Since many sampled recordings consist of calls from multiple frog species, frog call classification can be framed as a ML learning problem. However, previous studies have not adopted ML learning to classify frog calls. Therefore, it is worth investigating different ML learning algorithms for the classification of multiple vocalising frog species in field recordings for its scalability and flexibility [Read et al., 2011]. The principle of the BR method is to solve a multi-label classification problem using multiple binary classifiers respectively. Similar to our previous work [Zhang et al., 2016], three classic single-label learning algorithms: decision tree, and multi-layer perceptron, and random forest. Here, random forest is selected as the base classifier, since our previous study of classifying frog species has already demonstrated its comparable performance [Xie et al., 2016].

6.3 Experiment results

Each 10-second recording is divided into frames of 512 samples and 50% frame overlap for STFT. For MFCCs and AWSCCs, window size and overlap are 512 samples and 50%, the window function is a Hamming window. All algorithms were programmed in Matlab 2014b except ML learning, which was implemented in Meka 1.7.7[4].

6.3.1 Evaluation metrics

For multi-label classification, the performance evaluation differs from the classic single-label classification systems. The multi-label classification results often have a situation where partial labels are correctly predicted, but the prediction of single-label classification is either correct or incorrect. Therefore, some traditional evaluation metrics for the single-label classification, such as precision, recall, and accuracy, are no longer suitable for the multi-label classification system. In this study, three evaluation metrics, hamming loss, accuracy, and subset accuracy, are used, where all the three are example based measures [Madjarov et al., 2012b].

The definition of Hamming loss can be found in Chapter 5.

Accuracy for a single instance x_i is defined by the Jaccard similarity coefficients between

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

the ground truth y_i and the prediction $h(x_i)$. Accuracy is micro-averaged across all examples:

$$accuracy = \frac{1}{N}\sum_{i=1}^{N} |\frac{h(x_i \cap y_i)}{h(x_i \cup y_i}| \tag{6.8}$$

where N is the number of instances, y_i denotes the ground truth of instance x_i, and $h(x_i)$ denotes the predictions for the same instance.

Subset accuracy is defined as follows:

$$subsetAccuracy = \frac{1}{N}\sum_{i=1}^{N} I(h(x_i) = y_i) \tag{6.9}$$

where $I(true) = 1$ and $I(false) = 0$. This is a very strict evaluation measure as it requires the predicted set of labels to be an exact match of the true set of labels.

The values for hamming loss, accuracy, and subset accuracy range from 0 to 1. For hamming loss, 0 denotes the perfect result, and 1 means the wrong prediction of all labels over every instance, whereas for accuracy and subset accuracy, the values have the completely opposite meanings.

6.3.2 Classification results

Experiment results are shown in Table 6.1. The combination of LPCs, MFCCs, and AWSCCs, can achieve the best performance with ML-random forest. Using only cepstral coefficients, the best classification performance is achieve by AWSCCs(2), of which the accuracy is 0.662. After adding the temporal feature (LPCs), the accuracy is improved to 0.694. The reason for this improvement might be that this combination can achieve both temporal and frequency information of the recording. Compared to other classifiers, the classification performance of ML-RF is better than both ML-kNN and ML-DT (Table 6.2).

6.3.3 Comparison with MIML

In this chapter, ML learning is used to classify frog calls without syllable segmentation. Compared to the MIML learning (Figure 5.4 in Chapter 5), the ML classification has a slightly better classification performance. For ML classification, LPCs, MFCCs, and AWSCCs are combined

Table 6.1: Comparison of different feature sets for ML classification. Here, MFCCs-1 and MFCCs-2 denote cepstral features are calculated via first and second methods, respectively

Feature set	Base Classifier	Hamming loss ↓	Accuracy ↑	Subset accuracy ↑
MFCCs-1	Random forest	0.143 ± 0.015	0.64 ± 0.035	0.354 ± 0.052
MFCCs-2		0.139 ± 0.01	0.659 ± 0.019	0.362 ± 0.036
AWSCCs-1		0.139 ± 0.011	0.656 ± 0.009	0.371 ± 0.038
AWSCCs-2		0.138 ± 0.012	0.662 ± 0.01	0.38 ± 0.031
AWSCCs-2 + LPCs		$\mathbf{0.117 \pm 0.011}$	$\mathbf{0.694 \pm 0.027}$	$\mathbf{0.424 \pm 0.027}$

Table 6.2: Comparison of different ML classifiers

Feature set	Base classifier	Hamming loss ↓	Accuracy ↑	Subset accuracy ↑
AWSCCs-2 + LPCs	ML-kNN	0.139 ± 0.023	0.679 ± 0.039	0.362 ± 0.065
AWSCCs-2 + LPCs	ML-Decision Tree	0.171 ± 0.012	0.593 ± 0.018	0.272 ± 0.036
AWSCCs-2 + LPCs	ML-Random forest	0.117 ± 0.011	0.694 ± 0.027	0.424 ± 0.027

for the classification. MFCCs and AWSCCs are calculated by averaging cepstral features in the temporal direction. Although this process will compress the information in the temporal direction, the information of the cepstral domain is obtained. Since most frog species tend to continuously make calls, the compression of the temporal information will not greatly affect the discriminability of the ceptral features. LPCs are added to get the temporal information. In contrast, features used for MIML classification are calculated from each segmented syllable. However, current AED often cannot accurately segment frog calls with low energies, which greatly affects the classification performance.

6.4 Summary

In this chapter, ML learning is used to classify multiple simultaneously vocalising frog species in field recordings. A combination of AWSCCs-1 and LPCs can achieve the best classification performance, which is similar with MIML classification results reported in Chapter 5. To construct the final feature set, a novel method is proposed based on spectral clustering. Although the classification performance is similar with the statistical method, the feature dimension is greatly decreased from 72 to 24. Compared to MIML learning, ML learning does not need to segment frog syllables, which can increase the robustness of the classification results. Compared to other classifiers, ML-random forest can achieve the best classification performance with a hamming

loss of 0.117.

Chapter 7

Conclusion and future work

This thesis has addressed frog call classification using both trophy and field recordings. For trophy recordings, a combined feature set using temporal, perceptual and cepstral features is proposed. A novel cepstral feature with good anti-noise performance is proposed using wavelet packet decomposition (WPD). To classify multiple simultaneously vocalising frog species in field recordings, two classification frameworks are adopted: multiple-instance multiple-label (MIML) learning and multiple-label (ML) learning.

Challenges of this thesis lie in designing effective feature extraction algorithms and adopting classification frameworks. Key contributions of this research to the challenges are summarised, and useful avenues of inquiry for improving the methods described in this thesis are explored.

7.1 Summary of contributions

In Table 7.1, the proposed algorithm of each chapter is listed.

Table 7.1: The list of algorithms used in this thesis

Algorithm ID	Data	Segmentation	Fature	Classifier	Contribution	Chapter
1	Trophy recordings	Amplitude frequency information	Various	kNN, SVM, RF, and NN	feature and integration	3
2	Trophy and field recordings	Amplitude frequency information	AWSCCs	kNN and SVM	feature	4
3	Field recordings	AED	Various	MIML-kNN, MIML-SVM, and MIML-RBF	segmentation and integration	5
4	Field recordings	No segmentation	LPCs, MFCCs, and AWSCCs	ML-kNN, ML-DT, and ML-RF	feature and integration	6

Detailed contributions of this thesis are summarised below:

(1) *An enhanced acoustic feature set for frog call classification in trophy recordings.*
Effectively modelling frog vocalisations has significant impact on the performance of frog call classification systems. A novel feature set is proposed to represent frog calls using temporal, perceptual, and cepstral information. A combination of temporal, perceptual, and cepstral features can greatly increase the discriminability of the combined feature set. Evaluations of the propose feature set are based on 24 frog species from trophy recordings. Five machine learning algorithms are compared to the proposed feature set. Background noise with SNR from -10 dB to 40 dB is added to test the anti-noise ability of the proposed feature set. Experimental results show that (1) Compared to previous feature sets, an enhanced feature set including 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 feature is very sensitive to the background noise, but can achieve high classification accuracy for high SNR recordings.

(2) *A novel feature via adaptive WPD for frog call classification in both trophy and field recordings.*
Cepstral features are widely used for classifying frog calls. Although cepstral features have shown high classification performance for classifying frog species in trophy recordings, the performance is quickly decreased when classifying frog species in field recordings. A novel cepstral feature via WPD is proposed to increase the anti-noise ability. An adaptive frequency scale is generated by applying k-means clustering to all dominant frequencies of training datasets. Compared to other frequency scales, the adaptive frequency scale can better reflect the frequency distribution of frog calls. Evaluations of the propose feature set are based on 18 frog species from trophy recordings and eight frog species from field recordings. Experimental results in both trophy and field recordings show that the propose cepstral feature can achieve the best classification performance when compared to other cepstral features using trophy recordings. For field recordings, the classification performance of our cepstral features does not greatly decrease unit the SNR is 10 dB.

(3) *Design a MIML classification framework for frog call classification in field recordings.*
Most field recordings contain multiple simultaneously vocalising frog species, a single-instance single-label (SISL) classification framework might be unfit for classifying frog species in those field recordings. Compared to SISL learning, MIML learning is a natural fit for field recordings

of frogs. A novel MIML classification framework is adopted to focus on frog calls. To segment individual frog syllables, a novel AED algorithm is designed based on event filtering. The propose classification framework is evaluated using 342 10-second recordings including eight frog species. Experimental results show that MIML-RBF achieves the best classification results with shape based feature sets. Compared to SISL learning, MIML learning can significantly increase the classification results for all eight frog species.

(4) *Design a ML classification framework for long-term monitoring of frogs in field recordings.*
Compared to SISL learning, MIML learning shows a better performance for classifying multiple simultaneously vocalising frog species in field recordings. However, MIML classification performance is highly affected by the AED method. To reduce the effect of AED, a novel ML classification framework is adopted. A new method for constructing cepstral features is proposed. Evaluations of the ML classification framework are based on the same dataset with MIML learning. The ML classification performance is slightly better than MIML learning, but non-use of the segmentation process can greatly increase the classification efficiency.

7.2 Limitations and future work

Although our proposed frog call classification framework shows promising classification performance, there is still much work that can be done to help scientists and researchers in data collection and analysis of the bio-acoustics communities.

- For one frog species, calling parameters of different areas might have some variations. It is necessary to investigate our proposed classification framework for classifying frog vocalisations from different areas.

- Since field recordings often contain much background noise, it is important to develop effective noise reduction algorithms to reduce the background noise and improve the classification performance.

- For each frog, there are many types of frog calls: (1) mating calls, (2) territorial calls, (3) male release calls, (4) female release calls, (5) distress calls, and (6) warning calls. Among them, almost all studies that use machine learning algorithm to classify frog calls

select mating calls (advertisement calls) as the research targets. It it worthwhile to classify frog calls into different species and further classify different types of calls for one frog species.

- One aspect that requires further 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 field recordings often have many simultaneous overlapping calling activities from birds, frogs, insects, and many other sources.

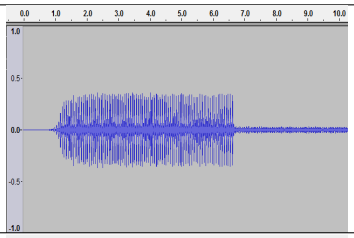
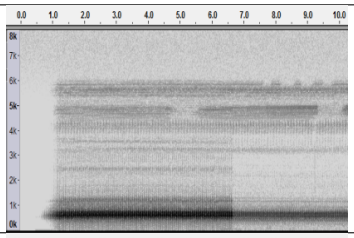
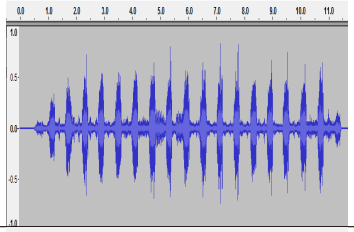
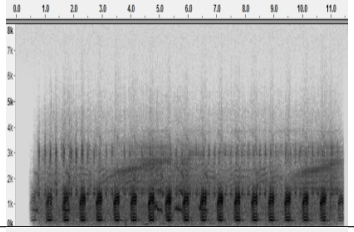
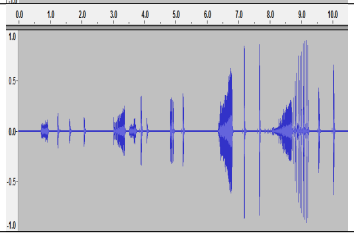
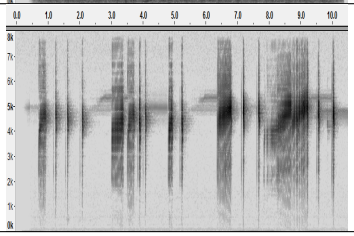
- Our developed frog call classification framework aims to help ecologists to study frogs over larger spatial and temporal scales. However, there is still no a generic platform for running frog recordings. It is necessary to develop a toolbox with an easy user interface for frog call classification, and then ecologists can conduct the analysis on their own. We focus on efficacy in this research, however efficiency is also very important in big data analysis. 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.

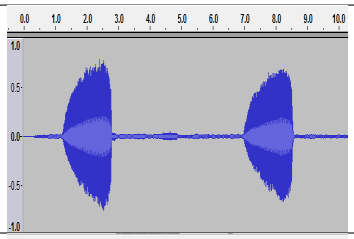
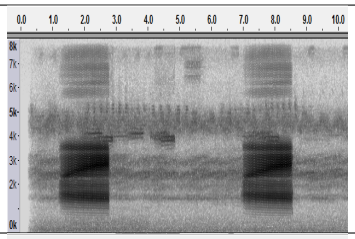
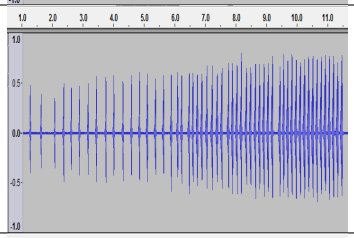
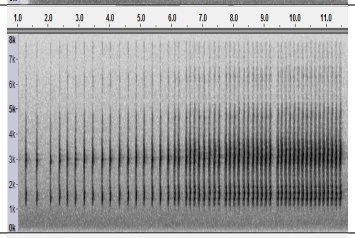
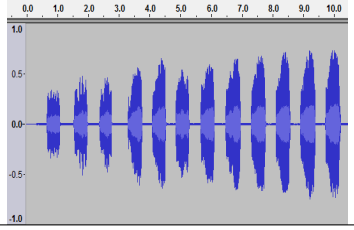
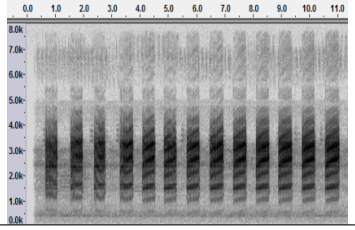
Appendix A

Waveform, spectrogram and SNR of frog species from trophy recordings

---

Table A.1: Waveform, spectrogram, and SNR of selected six frog species from trophy recordings. The SNR is calculated as $SNR = 10 \times log_{10}[(\sum_{i=m}^{m+L} S_i^2)/(\sum_{j=n}^{n+L} N_j^2)]$, where L is the length of the signal and noise used for calculating SNR, and set at 6000 samples, n and m are manually selected start location in the waveform for noise and signal, respectively

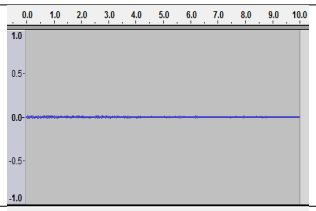
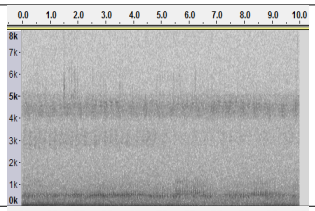
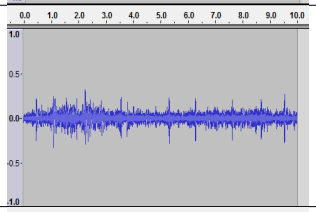
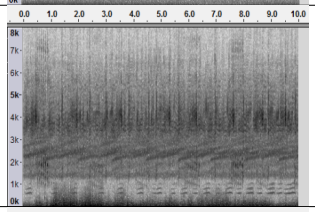
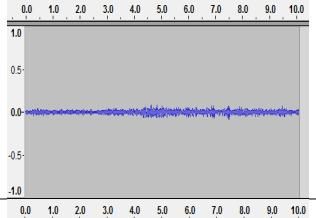
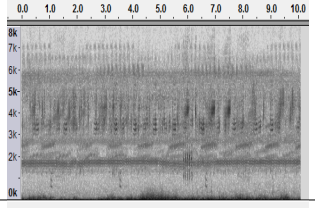
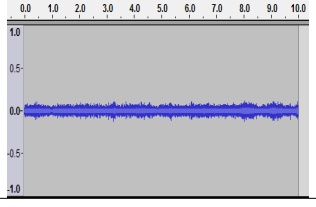
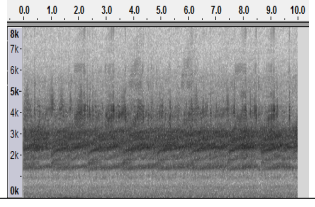
	Waveform	Spectrogram	SNR (dB)
Bufo marinus			19.35
Litoria caerulea			15.78
Litoria fallax			43.7

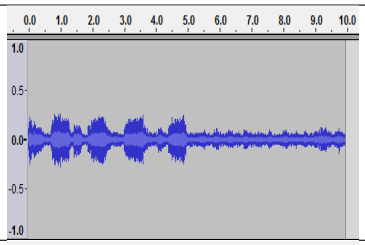
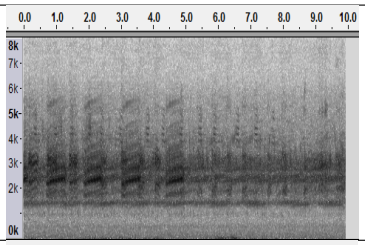
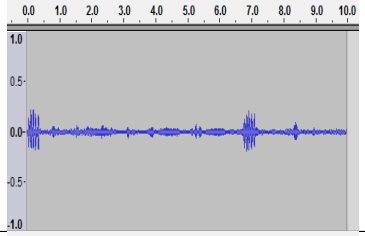
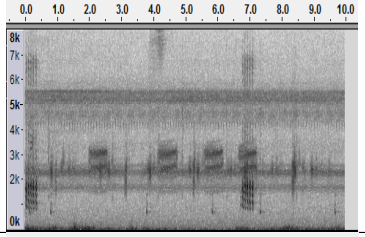
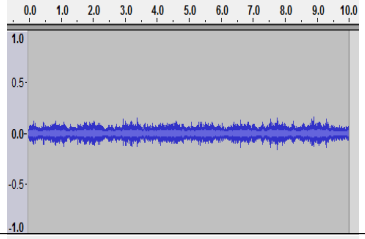
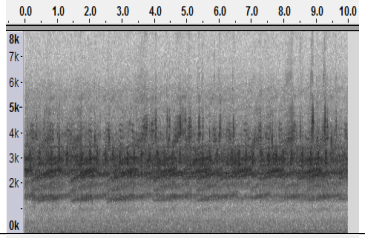
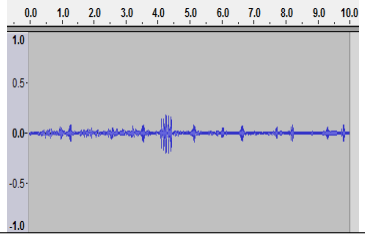
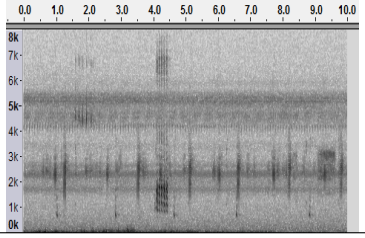
Litoria gracillenta			25.8
Litoria latopalmata			35.85
Litoria rubella			36.2

Appendix B

Waveform, spectrogram and SNR of six frog species from field recordings

Table B.1: Waveform, spectrogram, and SNR of eight frog species (field recordings)

	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

References

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.

Bao, L. and Cui, Y. (2005). Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics*, 21(10):2185–2190.

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.

Biswas, A., Sahu, P., and Chandra, M. (2014). Admissible wavelet packet features based on human inner ear frequency response for hindi consonant recognition. *Computers & Electrical Engineering*, 40(4):1111 – 1122.

Böll, S., Schmidt, B., Veith, M., Wagner, N., Rödder, D., Weinmann, C., Kirschey, T., and Loetters, S. (2013). Amphibians as indicators of changes in aquatic and terrestrial ecosystems following gm crop cultivation: a monitoring guideline. *BioRisk*, 8:39.

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.

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X. Z., Raich, R., Hadley, S. J., Hadley, A. S., and Betts, M. G. (2012). Acoustic classification of multiple simultaneous bird species:

A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131(6):4640–4650.

Camacho, A., García-Rodríguez, A., and Bolaños, F. (2013). 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.

Carey, C. and Alexander, M. A. (2003). Climate change and amphibian declines: is there a link? *Diversity and distributions*, 9(2):111–121.

Chen, W., Zhao, G., and Li, X. (2013). A novel approach based on ensemble learning to nips4b challenge. In *proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod. org/nips4b, joint to NIPS, Nevada*.

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. *International Science and Technology Conference*.

Colonna, J., Ribas, A., dos Santos, E., and Nakamura, E. (2012a). 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.

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

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.

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.

Duellman, W. E. and Trueb, L. (1994). *Biology of amphibians*. JHU press.

Dufour, O., Artieres, T., Glotin, H., and Giraudet, P. (2013a). Clusterized mel filter cepstral coefficients and support vector machines for bird song identification. In *In Proc of 1st workshop on Machine Learning for Bioacoustics*, volume 951, pages 89–93.

Dufour, O., Glotin, H., Artires, T., Bas, Y., and Giraudet, P. (2013b). Multi-instance multi-label acoustic classification of plurality of animals: birds, insects & amphibian. *Workshop on Neural Information Processing Scaled for Bioacoustics*, pages 164–174.

Farooq, O. and Datta, S. (2001). Mel filter-like admissible wavelet packet structure for speech recognition. *IEEE Signal Processing Letters*, 8(7):196–198.

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.

Glotin, H., LeCun, Y., Artieres, T., Mallat, S., Tchernichovski, O., and Halkias, X. (2013a). Neural information processing scaled for bioacoustics, from neurons to big data. usa (2013). *URL: http://sabiod. org/NIPS4B2013_book. pdf*.

Glotin, H., Sueur, J., Artières, T., Adam, O., and Razik, J. (2013b). Sparse coding for scaled bioacoustics: From humpback whale songs evolution to forest soundscape analyses. *The Journal of the Acoustical Society of America*, 133(5):3311–3311.

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.

Han, W., Chan, C.-F., Choy, C.-S., and Pun, K.-P. (2006). An efficient mfcc extraction method in speech recognition. In *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, pages 4–pp. IEEE.

Härmä, A. (2003). 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, pages V–545. IEEE.

Heller, J. R. and Pinezich, J. D. (2008). Automatic recognition of harmonic bird sounds using a frequency track extraction algorithm. *The Journal of the Acoustical Society of America*, 124(3):1830–1837.

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

Jaafar, H. and Ramli, D. A. (2015). Effect of natural background noise and man-made noise on automated frog calls identification system. *J. Trop. Resour. Sustain. Sci*, 3:208–213.

Jaafar, H., Ramli, D. A., and Shahrudin, S. (2013a). A comparative study of classification algorithms and feature extractions for frog identification system. *School of Electrical and Electronic 4th Postgraduate Colloquium*, 4.

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.

Jancovic, P. and Kokuer, M. (2015). Acoustic recognition of multiple bird species based on penalized maximum likelihood. *Signal Processing Letters, IEEE*, 22(10):1585–1589.

Juan Mayor, L. M. M. (2009). 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*.

Kular, D., Hollowood, K., Ommojaro, O., Smart, K., Bush, M., and Ribeiro, E. (2015). Classifying frog calls using gaussian mixture models. In *Advances in Visual Computing*, pages 347–354. Springer.

Lasseck, M. (2013.). Bird song classification in field recordings: winning solution for nips4b 2013 competition. In *Proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod. org/nips4b, joint to NIPS, Nevada.*

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.

Lei, B., Rahman, S. A., and Song, I. (2014). Content-based classification of breath sound with enhanced features. *Neurocomputing*, 141:139–147.

Litvin, Y. and Cohen, I. (2011). Single-channel source separation of audio signals using bark scale wavelet packet decomposition. *Journal of Signal Processing Systems*, 65(3):339–350.

Madjarov, G., Kocev, D., Gjorgjevikj, D., and D?eroski, S. (2012a). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084 – 3104.

Madjarov, G., Kocev, D., Gjorgjevikj, D., and Džeroski, S. (2012b). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104.

Mallawaarachchi, A., Ong, S., Chitre, M., and Taylor, E. (2008). Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles. *The Journal of the Acoustical Society of America*, 124(2):1159–1170.

Massaron, L. (2013). Ensemble logistic regression and gradient boosting classifiers for multilabel bird song classification in noise (nips4b challenge). *Proc of Neural Information Processing Scaled for Bioacoustics, joint to NIPS*.

Melter, R. A. (1987). Some characterizations of city block distance. *Pattern Recognition Letters*, 6(4):235 – 240.

Mencıa, E. L., Nam, J., and Lee, D.-H. (2013). Learning multi-labeled bioacoustic samples with an unsupervised feature learning approach. *Proc of Neural Information Processing Scaled for Bioacoustics, joint to NIPS*, 2013:184–189.

Mutschmann, F. (2015). Chytridiomycosis in amphibians. *Journal of Exotic Pet Medicine*, 24(3):276–282.

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

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

Razik, J., Hoeberechts, M., Doh, Y., et al. (2015). Sparse coding for efficient bioacoustic data mining: Preliminary application to analysis of whale songs. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 780–787. IEEE.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359.

Ren, Y., Johnson, M. T., and Tao, J. (2008). Perceptually motivated wavelet packet transform for bioacoustic signal enhancement. *The Journal of the Acoustical Society of America*, 124(1):316–327.

Roch, M. A., Brandes, T. S., Patel, B., Barkley, Y., Baumann-Pickering, S., and Soldevilla, M. S. (2011). Automated extraction of odontocete whistle contours. *The Journal of the Acoustical Society of America*, 130(4):2212–2223.

Ruiz-Munoz, J., Orozco-Alzate, M., and Castellanos-Dominguez, G. (2015). Multiple instance learning-based birdsong classification using unsupervised recording segmentation. *Proceedings of the Twenty-Fourth International Joint Conference On Artificial Intelligence (IJCAI 2015)*.

Selin, A., Turunen, J., and Tanttu, J. T. (2007). Wavelets in recognition of bird sounds. *EURASIP Journal on Applied Signal Processing*, 2007(1):141–141.

Somervuo, P. et al. (2004). Classification of the harmonic structure in bird vocalization. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 5, pages V–701. IEEE.

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

Stowell, D. and Plumbley, M. D. (2013). Feature design for multilabel bird song classification in noise (nips4b challenge). *Proceedings of NIPS4b: neural information processing scaled for bioacoustics, from neurons to big data*.

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*.

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

Towsey, M., Planitz, B., Nantes, A., Wimmer, J., and Roe, P. (2012). A toolbox for animal call recognition. *Bioacoustics*, 21(2):107–125.

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.

Xie, J., Towsey, M., Eichinski, P., Zhang, J., and Roe, P. (2015a). Acoustic feature extraction using perceptual wavelet packet decomposition for frog call classification. *e-Science (e-Science), 2015 IEEE 11th International Conference on*, pages 237–242.

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. *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4190–4194.

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). Acoustic classification of australian frogs based on enhanced features and machine learning algorithms. *Applied Acoustics*, 113:193–201.

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. (2013). Frog sound identification system for frog species recognition. In *Context-Aware Systems and Applications*, pages 41–50. Springer.

Zhang, L., Towsey, M., Xie, J., Zhang, J., and Roe, P. (2016). Using multi-label classification for acoustic pattern detection and assisting bird species surveys. *Applied Acoustics*, 110:91–98.

Zhang, X. and Li, Y. (2015). Adaptive energy detection for bird sound detection in complex environments. *Neurocomputing*, 155(0):108 – 116.

Zhou, Z.-H., Zhang, M.-L., Huang, S.-J., and Li, Y.-F. (2008). Miml: a framework for learning with ambiguous objects. *CORR abs/0808.3231*.