

AUTOMATIC FROG CALLS MONITORING:  
A MACHINE LEARNING APPROACH

By

QIANG FU

Bachelor of Science

Tsinghua University

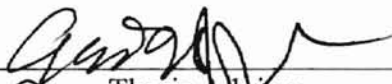
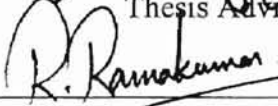

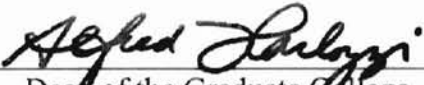
Beijing, P. R. China

1998

Submitted to the Faculty of the  
Graduate College of the  
Oklahoma State University  
in partial fulfillment of  
the requirements for  
the Degree of  
MASTER OF SCIENCE  
December, 2000

AUTOMATIC FROG CALLS MONITORING:  
A MACHINE LEARNING APPROACH

Thesis Approved:

  
\_\_\_\_\_  
Thesis Adviser  
  
\_\_\_\_\_  
  
\_\_\_\_\_  
  
\_\_\_\_\_  
Dean of the Graduate College

## PREFACE

Automatic recognition of frog vocalization is considered a valuable tool for a variety of biological research and environmental monitoring applications. This thesis proposes to develop an automatic, unattended monitoring system which can recognize the vocalizations of four species of frogs in the State of Oklahoma and can identify different individuals within the species of interest.

The proposed monitoring system deployed one directional microphone to record the frog calls in the field continuously. Sound signals were stored in digital audiotape first and then transmitted into a PC with WAVE file format. Species identification was performed first with the proposed filtering and grouping algorithm. Individual identification, which can detect different individual frogs within the same species, was performed in the second stage. Digital signal pre-processing, feature extraction, feature vector dimension reduction and pattern classification were performed step by step in this stage. Different feature extraction algorithms, including the time domain method (Linear Predictive Coding), the frequency domain method (Time-Dependent Fourier transform), and time-scale domain method (Wavelet Packet Transform), and two different dimension reduction algorithms are synergistically integrated to produce final feature vectors which were to be fed into a neural network classifier. The simulation results show the promising future of deploying an array of continuous, on-line environmental monitoring systems based upon nonintrusive analysis of animal calls.

## ACKNOWLEDGMENTS

I wish to express my sincere appreciation to my major advisor, Dr. Gary G. Yen for his intelligent supervision, constructive guidance, brainstorming, inspiration, and friendship. My sincere appreciation extends to my thesis committee members Dr. Rama G. Ramakumar and Dr. Jong-Moon Chung, whose guidance, assistance, encouragement, and friendship are invaluable.

I would like to thank the Department of Zoology and the Intelligent Systems and Control Laboratory at Oklahoma State University for supporting resources. I would also like to thank Dr. Paul Shipman for his technical support and providing all test data.

I would like to thank Dr. Bo Wang for his help in proof reading this thesis. I would also like to thank Kuo-Chung Lin, Phayung Meesad, and Zheng Wu for their previous work. I would also like to thank researchers in the Intelligent Systems and Control Laboratory for their discussions and recommendations.

Finally, I would like to thank my parents for their support and encouragement.

## TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION .....	1
1.1 Frog Call Monitoring System Overview.....	1
1.2 Problem Background.....	2
1.3 Motivation for the Research.....	4
1.4 Automatic Monitoring System Architecture.....	5
1.5 Thesis Outline .....	6
II. LITERATURE REVIEW .....	7
2.1 Overview of Animal Sound Recognition.....	7
2.2 Methods of Digital Signal Pre-processing .....	8
2.2.1 Pre-processing Case I.....	9
2.2.2 Pre-processing Case II.....	13
2.2.3 Pre-processing Case III .....	15
2.3 Methods of Pattern Classification .....	17
2.3.1 Case I: Statistical Pattern Classification.....	18
2.3.2 Case II: Neural Network Classification .....	19
2.3.3 Case III: Decision Tree Classification .....	21
III. FROG SPECIES IDENTIFICATION.....	24
3.1 Data Acquisition of Frog Call Signals .....	24
3.2 Spectrogram Analysis .....	25
3.3 Filtering and Grouping Algorithm .....	29
IV. IDENTIFICATION OF FROG INDIVIDUALS.....	40
4.1 Individual Identification Overview .....	40
4.2 Feature Extraction Algorithms.....	42
4.2.1 Linear Predictive Coding Method.....	42
4.2.2 Time-Dependent Fourier Transform Method.....	45
4.2.3 Wavelet Packet Transform Method.....	47
4.3 Dimension Reduction/Feature Selection Algorithm .....	52
4.4 Pattern Classification Algorithm.....	57
V. TEST RESULTS .....	59

Chapter	Page
5.1 Results for Species Identification.....	59
5.2 Results for Individual Identification .....	61
5.2.1 Data Segmentation .....	62
5.2.2 Generation of Training / Testing Data Set .....	62
5.2.3 System Description .....	63
5.2.4 Test Results .....	66
 VI. CONCLUSIONS AND FUTURE WORK .....	 71
6.1 Conclusions of the Research .....	71
6.2 Suggestions for Future Work .....	73
 REFERENCES .....	 75

## LIST OF TABLES

Table	Page
3-1 Latin Names, Roman Names and Abbreviation of Four Species of Frogs .....	25
3-2 Call Patterns of Different Species .....	28
5-1 Number of Individual Samples, and the Distribution of Training / Testing Data Sets .....	63
5-2 Test Results for LPC Method.....	67
5-3 Test Results for FT Method .....	67
5-4 Test Results for WPT Method.....	68

## LIST OF FIGURES

Figure	Page
1-1 Spectrogram of Southern Leopard Frog Call .....	3
1-2 Automatic Frog Call Monitoring System Architecture.....	5
2-1 Clear Call Spectrograms of 4 Species [38] .....	10
2-2 Unclear Call Spectrograms of 4 Species [38] .....	10
2-3 Frequency Track of a Call.....	11
2-4 A Sub Window Extracted From Figure 2-3 .....	12
2-5 Spectrogram of a Frog Chorus.....	16
2-6 Local Peaks of a Litroia Inermis Call.....	17
3-1 Spectrograms of 4 Different Species.....	27
3-2 Spectrogram of RAUT Call.....	29
3-3 One Period of Original Sound Signals $Y(t)$ .....	30
3-4 Frequency Response of the Designed Chebyshev Type II BPF .....	33
3-5 $Y_{bp}(t)$ , $Y_{sqr}(t)$ and $Y_{thres}(t)$ (from left to right) .....	34
3-6 Two Intervals Shown by Grouping Algorithm .....	36
3-7 Spectrograms of Two Typical Species [16] .....	37
3-8 Results of Continuously Calculation of 512-point FFT within One Data File .....	39
4-1 Filtered Signals (left) and Extracted Single Pitch (right).....	41
4-2 Wavelet Analysis Versus Time Domain, Frequency Domain and TDFT (Time- Dependent Fourier Transform) Analysis.....	47



Figure	Page
4-3 db8 Wavelet.....	48
4-4 Illustration of DWT of Signals.....	49
4-5 Three Level Wavelet Packet Decomposition Tree.....	50
4-6 Node Representation of WPD Tree.....	51
4-7 PDF of (a) Two Well-separated Classes and (b) Two Overlapping Classes .....	53
5-1 The Relationship Between LPC Coefficient Number $p$ and MSE.....	64

# CHAPTER I

## INTRODUCTION

### 1.1 Frog Call Monitoring System Overview

Recently there is an increasing interest and expenditure in environmental monitoring, both in North America and around the world. It is becoming essential to predict and assess the environmental impact of human activities on plants and animals. The populations of certain kinds of animals like birds and frogs are excellent indicators of overall environmental health. As many of the animals in an area may be heard but not seen, it is convenient to rely on their sounds as a means of identification. In many places manual census is not available, if not completely impossible. As a result, automatic recognition of animal sounds is then considered a valuable tool for biological research and environmental monitoring applications. In the present thesis work, an automatic, unattended, monitoring system is proposed to recognize the vocalization of different species of frogs in the State of Oklahoma.

The monitoring system, which does not require an expert attendance, deploys a directional microphone to capture frog calls continuously in the field, records sound signals into digital audio tapes, and translates them into digital audio data files. Species identification (including the use of different band pass filters and grouping algorithms) and individual identification within some species (including data preprocessing, feature selection and feature vector dimension reduction, and pattern classification) can be

automatically carried out in this monitoring system. Useful information such as the number of species identified and their approximate estimated population are then transmitted via Mesonet for follow-up environmental decision-making.

The successful development of this automatic monitoring system will provide a robust measurement to quantify environmental pollution. This system will greatly facilitate research to monitor the amphibian population as an indicator for environmental and water quality [1].

## **1.2 Problem Background**

Frog is a small, tailless animal with a squat body and long, powerful hind legs adapted for jumping. Most frogs have moist skin. They typically live both on land and in water. Toads are very similar to frogs except that toads typically have rough and dry skin and they often live in drier habitats. Frogs and toads are commonly acknowledged as the major divisions in amphibians. Amphibians lead a double life, alternately on land and in water. Typical amphibians include toads, newts, and salamanders as well as frogs. They usually live in temporary or permanent wetland areas such as woodland ponds and flooded fields.

Frogs are of great importance to humans. They are carnivores and consume large quantities of insects, worms or other small creatures. In turn, they may be a food source for other animals such as snakes. Frogs and toads are integral parts of the food chain. Many researchers in different fields are interested in frogs and toads because they are

considered to be bioindicators. The health of frog population is thought to reflect the health of the whole ecosystem. Auditory recognition of frogs is one feasible way to estimate frog population in the area of interest.

In the State of Oklahoma, different species of frogs and toads make calls starting from January and February till September. Frogs, as well as birds and whales, have developed the use of sound as the principal means of distant communication. Most species of frogs can produce two types of calls, a distress call and an advertisement call. Both males and females can make distress calls when they are in danger. Only males can produce advertisement calls, which are used to convey such information as location and breeding readiness to both sexes. Advertisement calls can be used to identify the species of frogs and toads in the State of Oklahoma.

After some practice a person with a "good" ear can easily learn to distinguish the calls of most species. For those with a less discriminating ear, an analytical graph of the sound can be a useful teaching aid for learning calls and comparing them. A spectrogram is a graphic image of sound that can be produced by digitizing sound in a computer and then showing it on a monitor. The spectrogram represents a plot of frequency against time; it depicts changes in frequency over the time duration of the call. Figure 1-1 shows a spectrogram of Southern Leopard frog call (three calls within the frame).

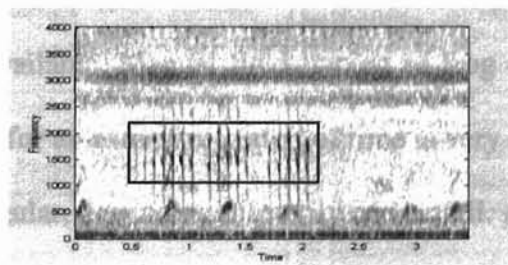


Figure 1-1: Spectrogram of Southern Leopard Frog Call

### 1.3 Motivation for the Research

Since early 1980, scientists have reported startling declines in the populations of some species of frogs [27]. These declines have occurred globally. Although the reason for frog population decline remains a mystery, there arise a variety of popular hypotheses and possible justifications, such as fluctuations caused partly by climatological changes, increased ultraviolet light due to anthropogenically caused ozone depletion, diseases and introduction of exotic species (e.g., bullfrogs). All of them make frog population an excellent indicator of environmental health, particularly in aquatic habitats because of their biphasic (aquatic and terrestrial) life. Also, frogs are in general susceptible to environmental toxicants due to their permeable skins. Understanding the effects of pollutants on frogs may help us to learn more about maintaining other species in their natural environments and about preserving these environments for human beings. Several species of frogs found in the State of Oklahoma are considered to be endangered in the foreseeable future (e.g., the leopard frog and the cricket frog).

Causes for frog population declines remain shrouded in mystery despite increased worldwide research efforts. Because of the difficulty and expensiveness of the censusing population of specific frog species, a conclusive analysis based on the estimation of frog population is not yet available. Manual field tracing of frog calls in extremely hot and high humidity wetlands for an extensive period of time is very difficult. And the activities of most species are irregular, depending primarily on rainfall. As a result, short field trips to these areas are not a reliable method to census the frog populations [1]. Therefore, an

automatic monitoring system, which remotely monitors calling anurans in the harsh environment, needs to be established with in-place Mesonet infrastructure.

#### 1.4 Automatic Monitoring System Architecture

Basically, the automatic frog call monitoring system can be functionally divided into two specific parts: species identification, followed by individual identification for certain species. First we collect sound signals by microphone, store them into digital audiotapes, and then transmit them in the form of digital WAVE files, which can then be fully analyzed by a digital computer. In species identification, filtering and grouping algorithm will be used. The individual identification includes three separate steps: signal preprocessing, possible feature vector dimension reduction, and pattern classification. Finally, useful information about frog species represented and the number of calls within certain time interval will be transmitted through Mesonet. Figure 1-2 illustrates the architecture of the system envisioned.

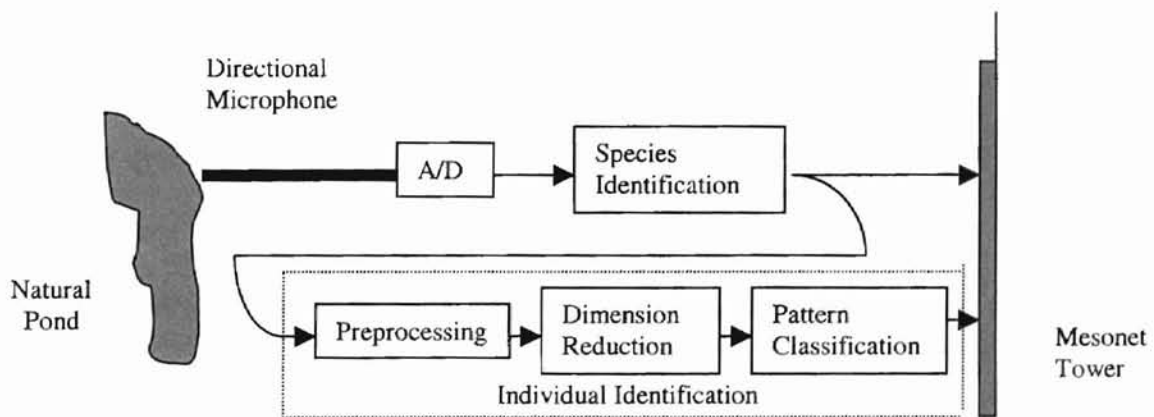


Figure 1-2: Automatic Frog Call Monitoring System Architecture

By establishing such an automatic system, which may sustain the harsh environment in the field, it has become possible to continuously monitor the number of different species of frog calls within the State of Oklahoma and approximately estimate the number of individual frogs within specific species of interest.

### **1.5 Thesis Outline**

In this thesis, Chapter 2 provides a literature review for different existing animal sound identification systems. Chapter 3 presents the proposed method of species identification, which is proven to be efficient and different from those mentioned in Chapter 2. Chapter 4 shows how individual identification works, which includes three major parts: signal preprocessing, feature vector dimension reduction, and pattern classification. Chapter 5 discusses the simulation results of species and individual identifications based on the available data sets. Chapter 6 gives the conclusion of the research and suggestions for future work.

## CHAPTER II

### LITERATURE REVIEW

#### 2.1 Overview of Animal Sound Recognition

Automatic recognition of animal vocalizations has been considered a valuable tool for a variety of biological research and environmental monitoring applications. Recently conducted researches have focused on the area of marine mammal population estimation [5, 22, 30] like whales and dolphins and species identification of groups such as birds [17, 18, 23, 35, 38] and frogs [39], which are frequently used as general indicators of diversity or ecological changes. Most of these applications focused on species identification, i.e., to identify different species of birds according to recorded birdsongs. However, only a few research efforts have been dedicated to quantify the repertoire of a single species and thus estimating population within the same species [5].

Manual censuses of animal vocalizations are often time-consuming, inaccurate, expensive and prone to observer biases. In some areas there are very few or even no suitably skilled observers. Automatic methods to identify and count animal vocalizations would allow a more extensive, more consistent and cheaper population monitoring for many species. It also can make intelligent deductions, which are not feasible with human observers.

Based on different characteristics of various animal sounds, there are different methods to perform automatic or manual recognition. Basically all recognition systems



include two stages: digital signal pre-processing and pattern classification. These two stages will be separately discussed as follows.

## 2.2 Methods of Digital Signal Pre-processing

The purpose of digital signal pre-processing is to extract a temporal measurement which contains useful information from the original data. As is known, these large volumes of original data sets generally contain only sparse segments of useful calls. These calls often have weak signal strength and possibly buried in interference, and usually consist of somewhat similar noises from other animals and the environment. Through the use of pre-processing we can extract only those useful signals critical for pattern recognition usage.

As with human speech, animal sounds can be sensibly interpreted using a time-frequency representation method. Thus tools designed for human speech analysis are commonly used for animal sound classification purpose. Generally this includes time domain methods such as linear predictive coding [17, 35], frequency domain methods such as Fourier transform, time-frequency domain methods as time dependent Fourier transforms, spectrogram [38, 39], and time-scale domain methods such as wavelet transforms [12]. In addition, biologists have considered zero-crossing analysis, autocorrelation functions, cepstral analysis, power spectral density (Welch method) and Wigner-Ville transforms as tools for pre-processing of signals.

In comparison with the human speech recognition problem, animal sounds are usually simpler to recognize than speech utterances. Their recognition would be an easy problem if it was conducted under similar conditions to that of most successfully deployed speech recognition systems: a single cooperative individual close to the microphone in a quiet environment [3].

Unfortunately these animal sounds are usually recorded in a much noisy environment, which means that we must recognize simpler vocalizations under much more difficult conditions. That is the main difference between human speech recognition and animal sound recognition. Work in the former case focuses on the utterances, while the latter focuses on robustly handling the recording conditions.

The noisy nature of animal sound identification lead to one conclusion: all those signal processing methods mentioned above provide only the necessary tools for the process of signal pre-processing. There is still much work to be done to extract the useful messages (or features) from raw signals. Three examples will be used to illustrate the works involved.

### **2.2.1 Pre-processing Case I**

The use of a time-frequency space such as spectrogram as input to a detection algorithm is proven to be effective, at least in part because the hearing physiology of many animals is constructed to produce time-varying spectral estimates [28]. Taylor [38] used spectrogram as the basic approach to process different species of birdcall signals.

Figure 2-1 shows examples of call spectrograms of 4 different species. As shown in Figure 2-1, automatic identification would not be difficult if the calls were recorded in a controlled environment. Unfortunately this is not the case.

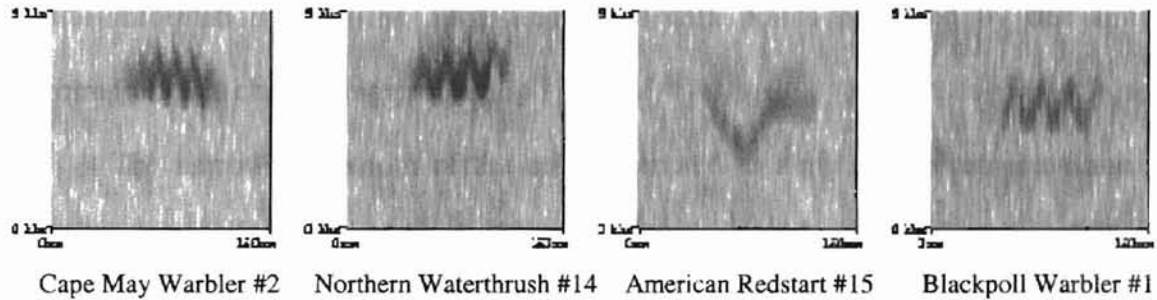


Figure 2-1: Clear Call Spectrograms of 4 Species [38]

In practice, birdcalls are recorded by specially designed microphones mounted on the roofs of buildings. Figure 2-2 displays examples of calls for the same 4 species as those in Figure 2-1 but demonstrating difficulties arising from the recording environment.

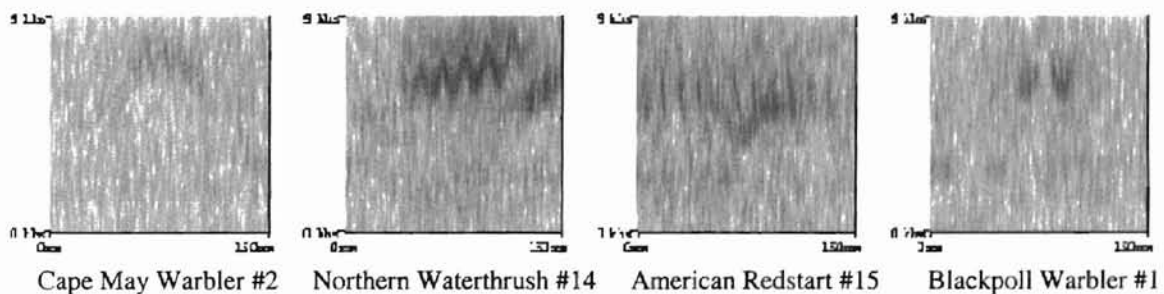


Figure 2-2: Unclear Call Spectrograms of 4 Species [38]

Each call in Figure 2-2 indicates a different problem. For example, the first one is very faint probably because it is far away from the microphone; the second call has a fainter call from an *Ovenbird* overlapping in time and close to it in frequency.

All these problems make the choice of representation very crucial. How to extract useful information from the huge amount of data points produced by one simple spectrogram is a challenging problem.

The narrow bandwidth of the calls prompted to simplify the representation by tracking the dominant frequency of the call. This reduces the two-dimensional call spectrogram to a one-dimensional track of the dominant frequency. Figure 2-3 gives an example of such a frequency track [38].

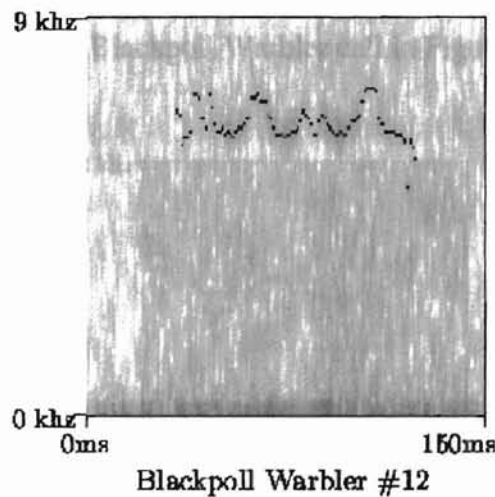


Figure 2-3: Frequency Track of a Call

Call frequency tracks are found by searching for sequences of local peaks in the spectrogram. Sequences, in which length or frequency changes out of the limits appropriate for calls, are rejected. Low energy peaks forming a spurious suffix or prefix

to a frequency track are detected by comparing the average energy of the peaks of the track. This is done by using an *ad-hoc* process [38].

In speech analysis, one method is to extract attributes from a fixed-sized window stepped through the signal. This method combined with machine learning has been applied very successfully to speaker recognition [36].

In this thesis, after some trial-and-error, a windowed approach to classify birdcalls is developed. A spectrogram is produced using a 64-point Discrete Fourier Transform (DFT) with a window size of 10ms, an increment of 3ms and a Hanning window.

A window of 11 peaks (totally 33ms) long slides along the frequency track. It is moved forward 1 peak at each step. According to the length of the call, 10 to 30 overlapping windows are produced. Figure 2-4 gives an example of such a window, which was extracted from the Blackpoll Warbler call in Figure 2-3.

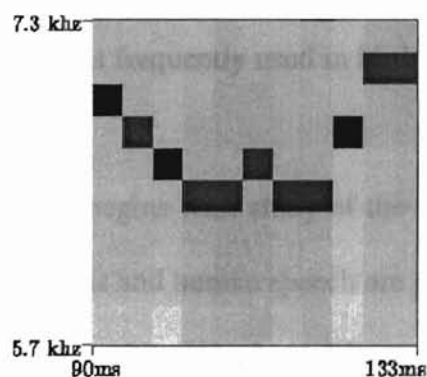


Figure 2-4 A Sub Window Extracted from Figure 2-3

Thirteen attributes were extracted from each window:

- The frequency of the peak at the center of the track.
- The frequencies of the other 10 peaks in the window.

- The energy of the spectrogram bins one and two increments above the central peak of the window, relative to the energy of the peak.

By this feature extraction method we obtained the final decision vector, which could be fed to the next stage: pattern recognition. With some modifications it finally correctly classified 79% of the 138 calls. Some other papers also used this method – spectrogram combined with local peak selection to extract feature vectors from original signal [39, 23].

### **2.2.2 Pre-processing Case II**

One of the most powerful speech analysis techniques, Linear Predictive Coding (LPC), which has been used successfully in speaker identification, in speech recognition and in speech compression, now is frequently used in birdcalls identification applications [17, 35].

Analysis of speech often begins with study of the vocal tract that created it. It is noticed that both bird vocalizations and human speech are generated by similar processes. If one assumes that vocalizations of both birds and humans can be modeled by source-filter interactions, one may then characterize these vocalizations by extracting filter coefficients from the time domain waveforms. LPC fits the impulse response of an all-pole filter to such a waveform. This suggests that LPC coefficients extracted from birdsong samples should retain enough information to permit identification of species.

One paper [17], which used LPC as pre-processing method to identify different bird species, will be focused in this section.

Comparing with spectrogram techniques used in previous example, the LPC method is much simpler and demands less computation. However, the identification system is not fully automatic. Data segments, which contain useful birdsong information, were left justified and picked out by hand first, then data were fed into the automatic pre-processing software package. The pre-processing stage includes several steps:

1. Framing – a non-overlapping Hamming Window, 256 samples (approx. 23.2ms) wide was applied. Frame widths for speech recognition are normally in the order of 10ms [13].
2. LPC – 15 time domain coefficients for a 15<sup>th</sup> order LPC filter were generated for each frame.
3. FFT – a fast Fourier transform of the 16 LPC coefficients was produced with 9 unique spectral magnitudes.

This procedure was repeated with a 1024 sample window for all songs. Initial investigation revealed that the overall length of a bird's song was an important cue in species identification. Therefore, a variable is introduced to represent the length of a song. The value of this variable was the same for each record within a given song. Spectral and time variables were normalized to a mean of zero and a standard deviation of one. Variables were squashed using a logistic function with a gain of unity.

Combined with pattern recognition method (which is a neural network in this case), the overall performance of attempting to identify 6 species ranged from 80% to

85% of correct identification. Except for some cases, mean performance increased as the original input data was changed progressively from 256 to 1024 and to combined data sets. This does not mean that the overall performance achieved with the 256 sample window was not good, however, increasing the window size and combining two resolutions of spectral data substantially improved the performance of the neural classifier, which means the second stage — pattern recognition in this kind of applications may play a more important role than pre-processing. Pattern recognition will be discussed in the next section.

### **2.2.3 Pre-processing Case III**

Very little research has been carried out in the field of frog call identification. Taylor [39] reported the development of a software system, which can recognize the vocalizations of 22 species of frogs that occur in an area of northern Australia. The system is based upon the classification of local peaks in the spectrogram of the audio signal using Quinlan's machine learning system, C4.5 [31]. Basically, the pre-processing method is similar to case I in Subsection 2.2.1.

The vocalizations of 22 frog species range in length from less than 20ms to over a second. Some species repeat their calls incessantly; other species usually make only occasional isolated calls. Many species tend to call in choruses with hundreds of individuals from a number of species present. The background noise includes that made by insects and rains. Some species' calls can be noise for others because of their



similarity. An example of the spectrogram of a frog chorus recorded can be seen in Figure 2-5.

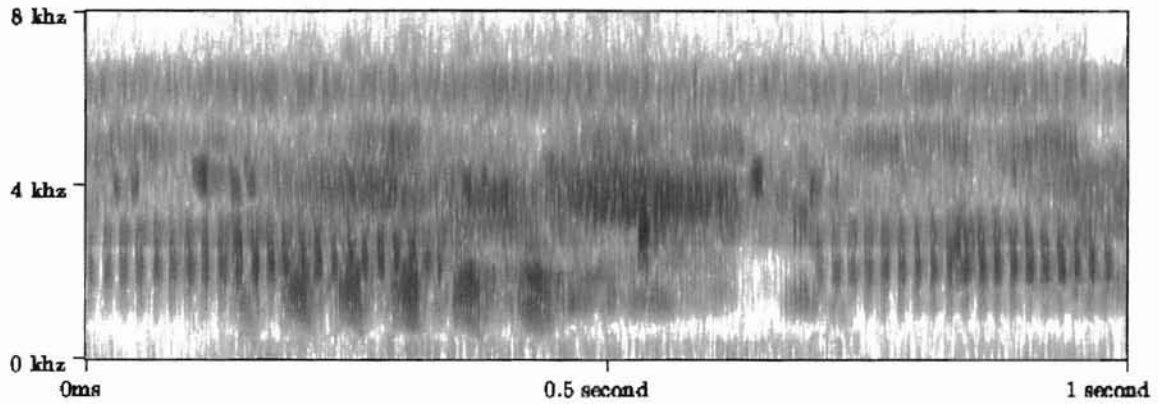


Figure 2-5: Spectrogram of a Frog Chorus

Figure 2-5, which is only a one second spectrogram, shows at least 11 individuals of 6 species of frogs and at least 3 species of insects, which can be considered as noise.

Most of the frog calls function as an advertisement to other members of the same species and hence have evolved to be species-specific. Experiments on other frog species have shown a variety of properties, which can be used by frog species to recognize their own species [6]. These include call rate, call duration, amplitude-time envelope, waveform periodicity, pulse-repetition rate, frequency modulation, frequency, and spectral patterns.

Because frog calls were collected in a noisy environment, the system makes no attempt to segment or isolate individual vocalizations. It works entirely from the spectrogram of the incoming audio signals. By means of the similar pre-processing method as discussed in Subsection 2.2.1, the system can pick out those local peaks appeared in the spectrogram. Figure 2-6 contains a call with the local peaks marked.

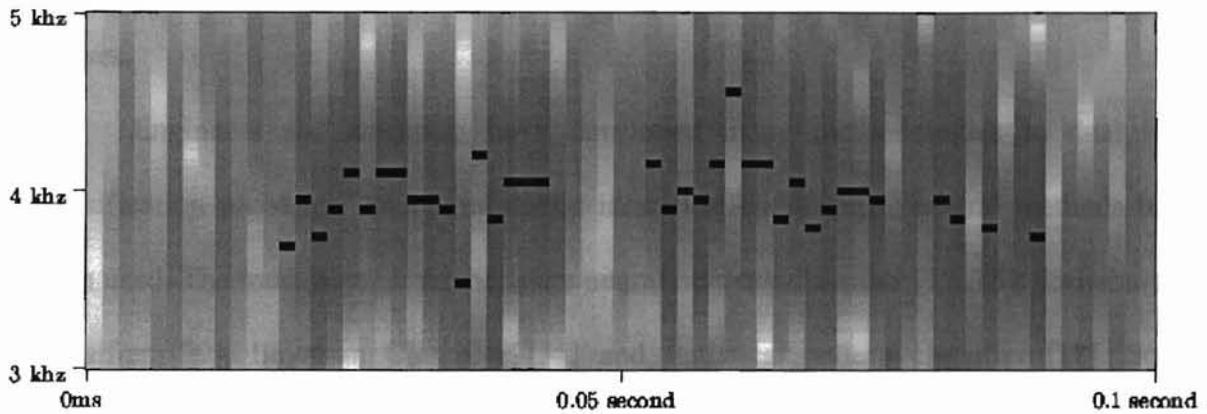


Figure 2-6: Local Peaks of a Litroia Inermis Call

The system examines each of the 40 local peaks in Figure 2-6 individually and classifies it as a particular species. Information from the spectrogram surrounding the peak is used to construct attributes for classification. These information include the frequency of the peak, the relative frequency of nearby peaks in preceding and succeeding time slices, and etc. There exist a significant number of ways that attributes constructed from this information. A set of approximately 70 possible attributes was constructed and a greedy search similar to the term *forward selection* [8] was used to choose a subset of 15 of the attributes that can be used for the system.

### 2.3 Methods of Pattern Classification

Pattern classification forms a fundamental solution to different problems in real world applications. The function of pattern classification is to categorize an unknown pattern into a distinct class based on a suitable similarity measure. Thus similar patterns

are assigned to the same classes while dissimilar patterns are classified into different classes.

Engineers and scientists have developed many methodologies to deal with classification problems. In animal sound identification systems, several methods have been used. The most popular methods are neural network classifier [17, 18], decision-tree classifier [39], Bayesian Classifier [35] and statistical pattern classifier [18]. Some researchers combined more than one method together in their applications. In the following, statistical pattern classifier, neural network classifier, and decision-tree classifier will be briefly discussed.

### **2.3.1 Case I: Statistical Pattern Classification**

As a traditional technique for classification problems, statistical pattern classification has been applied to numerous sound call recognition applications. This classical classification technique makes use of statistical decision theory to classify patterns.

In one research, a total of 133 birdsong records of 6 different species need to be classified [18]. After the pre-processing period, 23 variables were extracted from each birdsong signal. Then the number of variables was reduced from 23 to 8 using statistical analysis.

Preliminary examination of the correlation structure of the data indicated complex intercorrelation of variables. This in turn suggested that a smaller subset of variables

should be expected to contain enough information to permit discrimination. A stepwise discriminant analysis was performed. In this paper, separate ANOVA's (analysis of variance) and Scheffe's multiple comparison test [25] were calculated for each variable. The rest was checked for normality using the Shapiro-Wilk statistic  $W$  calculated and homoscedasticity with Bartlett's test [4] using a statistical SAS program (SAS is a kind of software used for statistical analysis.) [4].

Another way to explore the structure of data is to use ordination methods [29]. The amount of overlap among the songs of the 6 bird species was examined by reducing the eight variable data set to a two-dimensional (2-D) space using principal components analysis (PCA) and canonical discriminant analysis (CDA). The correlation option was used for PCA so that each variable was given equal weight in the analysis regardless of its scale. CDA was then applied to the reduced data set, with prior probabilities assumed to be equal. Preliminary tests indicated that quadratic discrimination functions were necessary since covariance matrices for different classes were too dissimilar to allow these classes gathering together. The results for quadratic discrimination analysis (QDA) were excellent. Overall accuracy was 93.3%, a bit better than the results using neural network method (neural network classifier was also used in pattern classification.).

### **2.3.2 Case II: Neural Network Classification**

Automatic pattern classification has been seriously considered by scientists and engineers from different fields. Many researchers have paid attention to neural network classifiers because of its capability of model-free and trainable systems, parallel

computation, and noise tolerance of neural networks. Basically, two properties of artificial neural network (ANN) inspire researchers to study neural network applications to deal with different pattern classification problems. They are:

1. The equivalent weighting matrix of ANN is determined by training, thus it can converge to a more optimal solution.
2. ANN is a kind of nonlinear estimator, which can embody more sophisticated responses.

In one study [18], by using LPC in pre-processing stage, each record composed of 10 variables derived from either the 512-sample window (9 variables) or from the 2048-sample window (9 variables) and the song length (1 variable). In this study, backpropagation was employed as the learning model [19]. Experiments with cross validation showed an ANN with 10 inputs, 12 hidden neurons and 6 outputs (10-12-6) can obtain the best result. The learning rate was set to 0.2. Target values of 0.0 and 1.0 were changed to 0.2 and 0.8 separately to accelerate learning [7]. Songs were divided into test and training sets in random order.

The six output values were analyzed for each of the test set records, and errors were computed. In this research, any output value greater than 0.6 for one of the six species was counted for that record. This was designed to yield consistency and somewhat conservative results.

The results from this ANN classifier were good. The overall performance ranged from 91% to 93% correct identification. But the drawback of this method was that it required considerable computation. Due to the dimensionality and the number of input

records generated during pre-processing, it takes several hours on a high-performance workstation to train the network. When larger sample sizes and more species are added, it may take even more time. For these reasons the statistical method is added as another classifier, which was mentioned in Subsection 2.3.1.

### 2.3.3 Case III: Decision Tree Classification

Until now, very few researchers have studied frog call identification. As mentioned above [39], by extracting about 70 possible attributes from the spectrogram, one feature selection algorithm similar to that in [8] was used and a subset of 15 of the attributes was obtained for the identification system.

Quinlan's machine learning system, C4.5 [31], was used to construct the classifier. C4.5 is a supervised learning system. It used a set of classified cases and a number of attributes for each case as training data and produced decision tree to classify further cases. The training data for C4.5 was extracted from calls of each 22 species. These calls were the recordings of single individuals with high quality. A number of them were manually chosen from each recording for training. This ensured that only calls from those required species were present in each piece of training data.

The decision tree produced by C4.5 has roughly 5,000 nodes. A small portion of the tree is shown as follows.

```
vert2 <= 18:
|   freq-4 <= -8: Uperoleia lithomoda
|   freq-4 > -8:
|   |   verta+vertb-vert <= 7: Litoria bicolor
|   |   verta+vertb-vert > 7:
|   |   |   timef+4 <= 0: Uperoleia lithomoda
```

```
| | | timef+4 > 0: Litoria caerulea
vert2 > 18 :
|   horiz <= -50: Litoria tornieri
|   horiz > -50: Uperoleia inundata
```

One C program developed in this research was used to implement this decision tree structure for identification purpose.

The identification of those local peaks is not very reliable. The error rate approaches 50%. Obviously this is unacceptable. Here, following this decision tree classifier the author developed one hierarchical structure of time segments based upon the typical temporal patterns of each frog species to carry on further identification. In the following, an example from [39] is cited for the completeness of the presentation.

“For example, a species might have a vocalization typically lasting 300ms containing a number of 30ms “notes” and it might usually produce 4 or more vocalizations in 3 seconds. Our system models this with 3 levels of segments. The level 0 segments will be 30ms long. If a threshold number of local peaks occur in that time period then the species is regarded as present in that level 0 segment, in other words we assume we have recognized a single “note” belonging to the species.

The level 1 segment will be 300ms long. If a threshold number of level 0 segments are identified as containing the species within that time period then the species is regarded as present in that level 1 segment, in other words we assume we have recognized a single vocalization of the species.

Similarly the level 2 segment will be 3s long and a threshold number of level 1 segments will be required to regard the species as present in the level 2 segment and hence reliably identified.”

Only a few species were required to use three level hierarchies. Most species only needed a one or two level hierarchy.

The performance of this system was good. Yet there were several mismatches of one species and one mismatch of a second species. Attempts to remedy this by modifying the temporal segments used for identifying these species were tried in this work. And for some species the system can only identify only one third of the total number of calls. This is partly because the system is following the rule that it would rather fail to recognize a call (a false negative) than to incorrectly indicate the presence of one species (a false positive). To further improve the prediction accuracy, there still remains appreciable works to be exerted.



## CHAPTER III

### FROG SPECIES IDENTIFICATION

#### 3.1 Data Acquisition of Frog Call Signals

The frog call monitoring system is based upon in-field acquisition of natural sound signals. One directional microphone, Telinga Pro V Mono Parabolic microphone was used to collect audio signals. It has frequency response: 40-18,000Hz: +/- 3dB. The microphone was connected with a SONY PCM-M1 digital audio recorder. Audio signals were stored in digital audiotape (DAT). Each DAT has capacity of 120 minutes. The Turtle Beach System, which includes Turtle Beach Montego II sound card mounted on PC, Turtle Beach AudioStation, Turtle Beach AudioView and additional supporting software, was utilized to transform signals stored in DAT into computer with digitized WAVE format. Each WAVE file has PCM (Pulse Code Modulation) format with 8,000Hz sampling frequency, 16-bit accuracy and monotony. Because no integrated hardware is available, all identification and classification were performed in lab rather than in field.

The capacity of each WAVE file is 16KByte per second. Considering about the computational complexity and the characteristics of frog calls, each WAVE file was chosen to be approximately 10-second long. To analyze data in real-time, the computer system should carry on species identification and the corresponding individual identification within 10 seconds. Later a Pentium III 500 PC was used for simulation

purpose. It was found that all numerical computations could be done within one or two seconds. That is well below 10 seconds. Hence a length of 10-second file segment is practically reasonable. After developing the whole hardware system, sound signals can be separated into 10-second long intervals automatically and be analyzed one by one in real-time.

### 3.2 Spectrogram Analysis

All frog calls acquisition works were performed in Stillwater, Oklahoma. Frog calls of four different species have been collected from January to May 2000. The names of each species are listed in Table 3-1. In the following the abbreviations of frogs of interest are being used for simplicity.

Table 3-1: Latin Names, Roman Names and Abbreviation of Four Species of Frogs

Subject Code	Latin Name	Roman Name	Abbreviation
01	<i>Rana utricularia</i>	Southern Leopard Frog	RAUT
02	<i>Bufo americanus</i>	American Toad	BUAM
03	<i>Pseudacris streckeri</i>	Streckeris Chorus Frog	PSST
04	<i>Pseudacris clarkii</i>	Spotted Chorus Frog	PSCL

Most of the frog calls function as an advertisement to other members of the same species and hence have evolved to be species-specific. Early there were experiments on

some species of frogs in other areas. These experiments have shown the fact that a variety of properties of frog calls can be used by one frog species to recognize the vocalizations of their own species [6]. To identify different species of frogs, some useful tools are used to explore properties of different frog calls. Spectrogram, which is frequently used in speech analysis, provides a three-dimensional representation of the sound intensity in different frequency bands over times and thus can be served as a powerful tool for the analysis purpose.

Spectrogram, essentially a type of time-dependent Fourier transforms, has two general classes: wideband spectrogram and narrowband spectrogram. A wideband spectrogram representation results from a window that is relatively short in time. It has poor resolution in frequency domain and good resolution in time domain. While a narrowband spectrogram uses a longer window to provide higher frequency resolution and with a corresponding decrease in time resolution [26]. In this research, single pitches of each frog calls need to be separated for identification and analysis purpose. These single pitches usually have very short durations (10-50ms). As a result, the wideband spectrogram, which has relatively better resolution in time domain, was used in the conducted research.

Figure 3-1 shows typical spectrograms of four different species being studied in this thesis. All unique call patterns are within those rectangle areas. One noticeable thing is that different noise sources seem to occupy nearly the entire frequency band.

Obviously each species has unique properties to distinguish themselves from each other. These include the call rate, the call duration, the amplitude-time envelope, the waveform periodicity, the pulse-repetition rate, the frequency modulation, the frequency,

and spectral patterns [6]. Since all frog call signals occupy frequency band well below 4KHz, according to sampling theory, the sampling rate ( $F_s$ ) of 8KHz is a reasonable choice. By using this sampling rate while not other higher sampling rate such as 11.5KHz, the system demands smaller memory and needs less computation time.

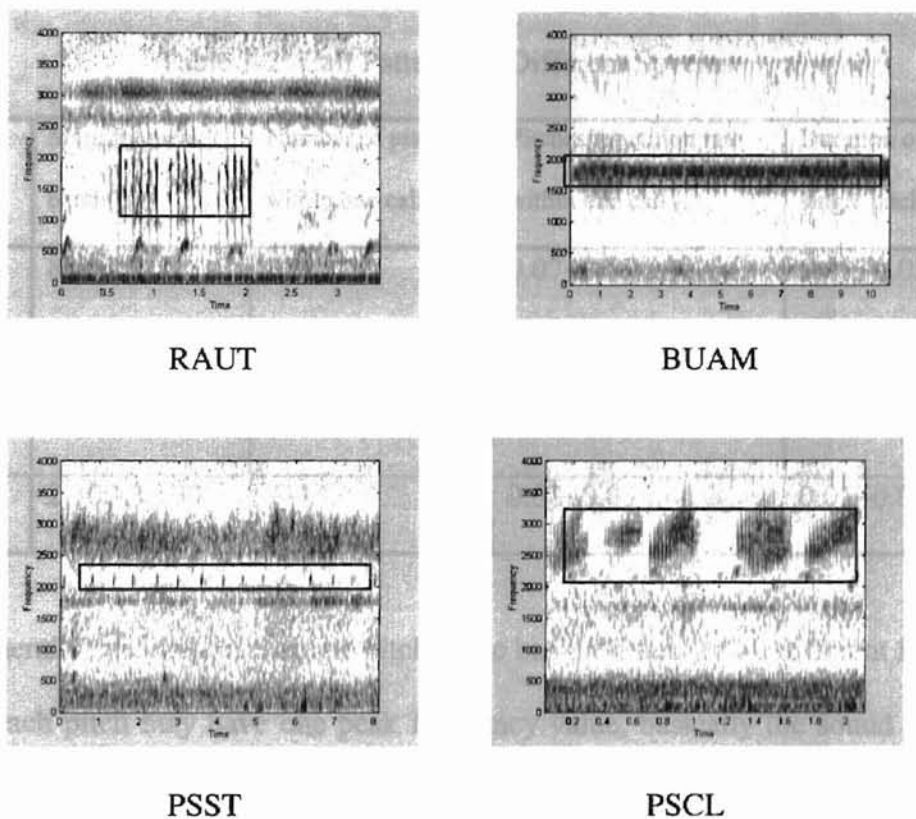


Figure 3-1: Spectrograms of 4 Different Species

These four species calls are all within certain frequency bands. According to call patterns they can be divided into three types. Type I includes RAUT and PSCL. Usually one frog of these types makes one to several calls at one time. Each call is composed of several pitches (pulses) with similar repetition rate. PSST can be seen as type II. Each call of PSST is only composed of one single pitch. Type II can be considered as the

simpler case of type I. Type III contains BUAM. This call usually lasts minutes, which can be regarded as continuous calls. Each call of this type is composed of many single pitches with similar repetition rate. Table 3.2 describes some typical features of these four different calls.

Table 3-2: Call Patterns of Different Species

Species	Main frequency band (Hz)	Number of pitches within one call	Pitches repetition rate within one call (s)	Duration of one single pitch (s)
RAUT	800-2,200	3~9	0.07~0.09	0.02~0.03
BUAM	1,600-2,000	Hundreds	0.04~0.05	0.03~0.04
PSST	2,000-2,500	1	N/A	0.04~0.05
PSCL	2,300-3,100	7~14	0.02~0.03	0.01~0.015

Generally speaking, one single pitch is the basic element of all different frog calls. Although each pitch may have one peak frequency value, basically the sound energy is evenly distributed within the whole frequency band. Also different individuals within the same species may have slightly different peak frequency value.

According to different call patterns shown in Figure 3-1, two general methods can be used to carry on species identification. One for Type I and II and the other for type III. Since RAUT and PSCL are similar and PSST is simpler than these two, methods proposed for detecting RAUT will be explained in detail as an example in the following section and method for identifying BUAM will also be illustrated.

Mithun Chandra Thirumala



as vertical striations in shaded area. The lengths of these two calls are  $T_1$  and  $T_2$  respectively. The pitch repetition rates within one call are roughly the same. That is,  $t_1 \approx t_2 \approx t_3$ .

After carefully studying all available RAUT call signals, the following conclusions were made.

1. All RAUT calls have mostly energy within 1,000-2,000Hz.
2. All RAUT calls are composed of several pitches.
3. The pitch repetition rate within one call is approximately 0.07-0.09s.
4. Single pitch length usually ranges from 0.02s to 0.03s.

Based on these observations, we can use one simple but efficient algorithm to identify RAUT calls from the whole data set. Below we use one period (roughly 3.5s) of original sound signals  $Y(t)$  as an example.

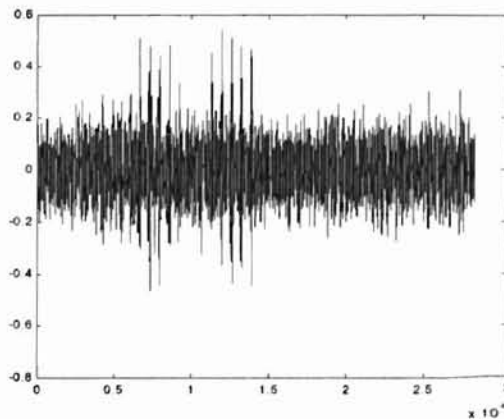


Figure 3-3: One Period of Original Sound Signals  $Y(t)$

Two calls can be roughly seen in this period, but they are not clear. Since RAUT calls are within certain frequency band, one bandpass filter (BPF) is introduced to filter other irrelevant signals out thus to give a more clear view.

Basically there are two kinds of digital filters, finite impulse response (FIR) filters and infinite impulse response (IIR) filters. Compare to FIR filters, IIR filters have the advantage that a variety of frequency-selective filters can be designed using the closed-form design formulas [26]. Once the problem has been specified with a given approximation method (e.g., Butterworth, Chebyshev, or elliptic), the order of the filter that will meet the specifications can be computed and the coefficients of the discrete-time filters can be obtained by straightforward substitution into a set of design equations. For the problem of interest, only magnitude response needs to be evaluated. There is no need to consider the frequency response. On the other hand, although the FIR filters have exact linear phase, there is no way to use any closed-form design equations. Also if we put aside phase considerations, it is generally true that a given magnitude response specification can be met most efficiently with an IIR filter. For these justifications IIR BPF is chosen in this case to filter out those irrelevant noises outside the specified frequency range.

There are mainly three kinds of IIR filters: Butterworth, Chebyshev and elliptic. The choice of the specific filter depends on the requirement of the filtered signal. Here we want to filter irrelevant signals out while keeping the magnitude of those signals within the passband unchanged. Since Chebyshev type II filters are monotonic in the passband and equiripple in the stopband, they can largely sustain the magnitude of signal components within the passband and attenuate unwanted signal components to the same level, no matter they are in high frequency band or low frequency band. For these reasons the Chebyshev type II filter was chosen as the desired BPF.



The desired passband is set from 1,000Hz to 2,000Hz. Listed below are parameters designed specially for this BPF [32]:

1. Passband corner frequency  $W_p$  :  $[1,050\text{Hz } 1,950\text{Hz}]/(F_s/2) = [0.2625 \ 0.4875]$ .
2. Stopband corner frequency  $W_s$  :  $[950\text{Hz } 2,050\text{Hz}]/(F_s/2) = [0.2375 \ 0.5125]$ .
3. Passband ripple  $R_p$  (the maximum permissible passband loss in decibels): 1dB.
4. Stopband attenuation  $R_s$  (the number of decibels the stopband is down from the passband): 50dB.

When choosing these parameters, the passband ripple  $R_p$  and stopband attenuation  $R_s$  are set to be 1dB and 50dB correspondingly. These are commonly used numbers that allow the system passes passband frequency components and mostly rejects all others. For digital filter design, the transition width ( $W_s - W_p$ ) is an important parameter. For large transition width value, the performance of this BPF may become worse. For small transition width value, the performance of BPF will be relatively good but as a result, the system may have impulse responses infinitely long. There exists one tradeoff when designing this parameter. Here the transition width is set to be 100Hz (0.025). In simulation process the BPF performs well while the order of this BPF is not very high.

Based on these parameters, the possible lowest order of Chebyshev type II filter is 12 and the Chebyshev type II cutoff frequency,  $W_n$ , that allows it to achieve the given specifications is  $[0.2472 \ 0.5089]$ .

The system equation of this Chebyshev type II BPF is

$$H(Z) = \frac{B(z)}{A(z)} = \frac{b(1) + b(2)z^{-1} + \dots + b(n+1)z^{-n}}{1 + a(2)z^{-1} + \dots + a(n+1)z^{-n}} \quad (3.1)$$

$$B = [b(1) \quad b(2) \quad \dots \quad b(25)] \quad (3.2)$$

$$A = [1 \quad a(2) \quad \dots \quad a(25)] \quad (3.3)$$

Through calculation, we derive the following filter coefficients:

$$B = [0.0080 \quad -0.0468 \quad 0.1482 \quad -0.3385 \quad 0.6258 \quad -0.9787 \quad 1.3242 \\ -1.5853 \quad 1.7154 \quad -1.7072 \quad 1.6081 \quad -1.5044 \quad 1.4630 \quad -1.5044 \\ 1.6081 \quad -1.7072 \quad 1.7154 \quad -1.5853 \quad 1.3242 \quad -0.9787 \quad 0.6258 \\ -0.3385 \quad 0.1482 \quad -0.0468 \quad 0.0080] \quad (3.4)$$

$$A = 1e+3 \times [0.0010 \quad -0.0071 \quad 0.0284 \quad -0.0814 \quad 0.1843 \\ -0.3456 \quad 0.5537 \quad -0.7721 \quad 0.9499 \quad -1.0401 \\ 1.0197 \quad -0.8986 \quad 0.7132 \quad -0.5101 \quad 0.3284 \\ -0.1898 \quad 0.0980 \quad -0.0449 \quad 0.0181 \quad -0.0063 \\ 0.0019 \quad -0.0005 \quad 0.0001 \quad -0.0000 \quad 0.0000] \quad (3.5)$$

The frequency response of this filter is shown in Figure 3-4. Magnitude response is shown on top and phase response is shown on bottom.

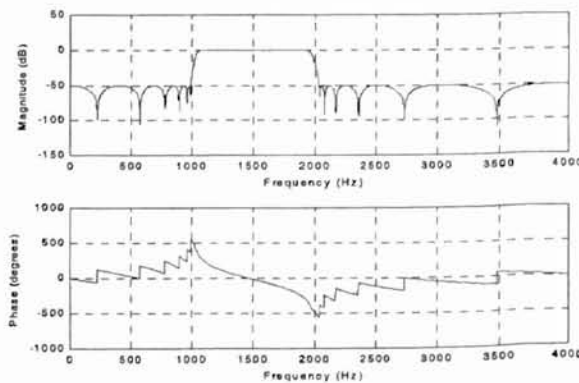


Figure 3-4: Frequency Response of the Designed Chebyshev Type II BPF

The filtered signal after this BPF is denoted by  $Y_{bp}(t)$ . Then  $Y_{bp}(t)$  is squared to get  $Y_{sqr}(t)$ . That is  $Y_{sqr}(t)$ , equals  $Y_{bp}^2(t)$ . Using the threshold those small values of  $Y_{sqr}(t)$  are zeroed out. The signal denoted by  $Y_{thres}(t)$  represents the signals after thresholding. These three steps of signal conditioning are shown in Figure 3-5.

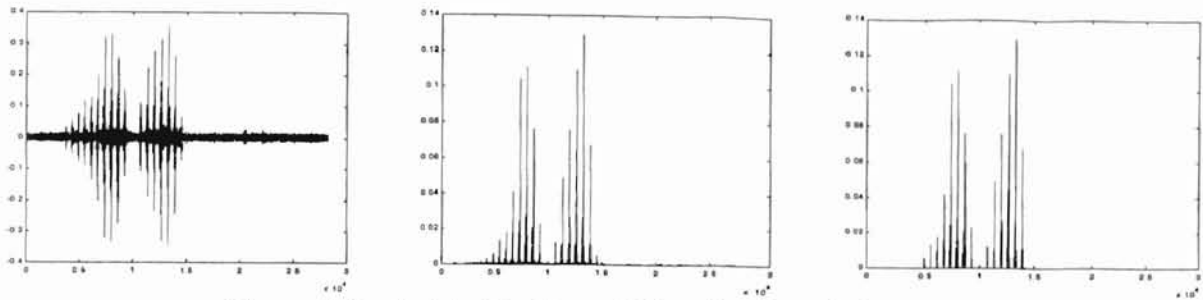


Figure 3-5:  $Y_{bp}(t)$ ,  $Y_{sqr}(t)$  and  $Y_{thres}(t)$  (from left to right)

The threshold,  $M$ , is set to be one positive number multiplying the mean value of signals  $Y_{sqr}(t)$ , which is proportional to the *average power* of  $Y_{bp}(t)$  [9]. In this case,  $M$  is  $10 \times \text{mean}(Y_{sqr}(t))$ . In  $Y_{thres}(t)$ , signals below this threshold (small spikes) are set to zero value thus to further reduce possible noise. From Figure 3-5, we can see that the filtered signals  $Y_{bp}(t)$  show two frog calls which are much more clear than the original signals. After thresholding, small noises are thrown away. Till now the first step has been done. Several sharp spikes can be seen in  $Y_{thres}(t)$ . But no judgment can be made that whether these spikes belong to true frog calls or not. Below one method called grouping algorithm is performed to obtain real RAUT frog call intervals and discard those false periods.

Single pitch of RAUT calls usually lasts 0.02s to 0.03s long and pitch repetition rate within one call ranges from 0.07s to 0.09s. The grouping algorithm is trying to

identify those spikes that tend to belong to one single call and group them together. Thus the time intervals of one single call can be detected. First by gathering those signals very close to each other together, the time intervals of those possible isolated spikes (pitches) are detected. Then by checking the duration of each short time interval, those with too short or too long intervals that obviously are not single pitches are thrown away. The second step of grouping is to group those pitches tend to belong to one call together by checking the intervals between adjacent pitches. Thus possible time intervals of single frog calls are acquired. Pitches too far away cannot be clustered together since they tend to belong to different calls. Finally the lengths of possible single calls are checked. Since one RAUT call is often composed of at least three strong isolated pitches, those intervals, which are not long enough, are discarded. After these three steps real RAUT call intervals are finally been detected and isolated. Figure 3-6 shows the result of two calls as well as several single pitches separated by the proposed grouping algorithm. The first call contains eight pitches. It starts from point  $t = 4,861$  and ends at point  $t = 9,261$ . So the duration of this call is  $0.55\text{s}$  ( $(9261-4861)/8000 = 0.55$ ). The second one contains six pitches. It starts from  $t = 10,686$  and ends at  $t = 14,013$ . The duration of this call is  $0.416\text{s}$ . The intervals of these single pitches are also been stored for possible further analysis like individual identification. The darker the shaded area, the stronger the single pitch is.

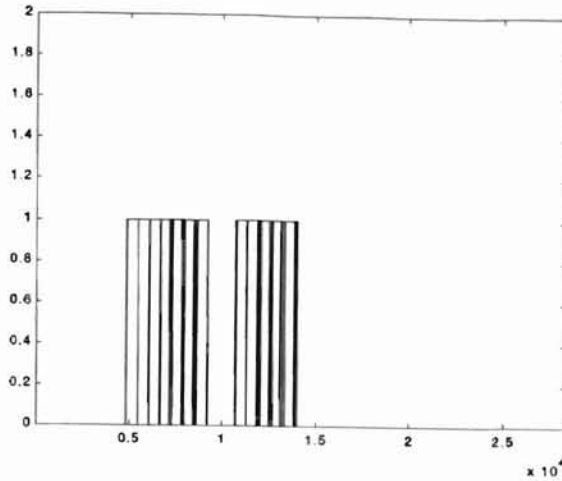


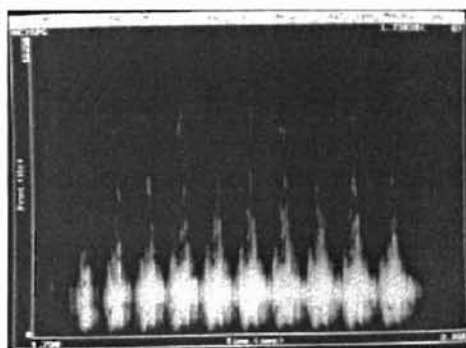
Figure 3-6: Two Intervals Detected by Grouping Algorithm

Compared to other species identification methods introduced in Chapter II; here we propose a simpler algorithm which requires less computation. For the incoming  $N$ -point data set, if the order of the proposed BPF is  $P$ , then  $4PN$  real multiplications and  $4PN$  real additions were required in filtering process;  $N$  real multiplications was required in squaring process;  $N$  real additions was required to calculate the mean value of  $Y_{sqr}(t)$ ; and approximately  $N$  comparisons was required in grouping algorithm. Thus the computational complexity of the proposed filtering and grouping algorithm is  $O(PN)$ . The computational complexity of spectrogram and Faster Fourier transform are both  $O(N \log_2 N)$  [26]. In this case  $P$  is 12 and  $N$  is 80,000. So  $\log_2 N$  is 16.3. The computation complexity of the proposed filtering and grouping algorithm and spectrogram/Fourier transform are in the same level. But all methods mentioned in Chapter II requires further processing of data with some pattern classification techniques, which require additional computation and time. The method proposed here doesn't need this step at all. So totally the proposed method needs less computation. The LPC method mentioned in Subsection

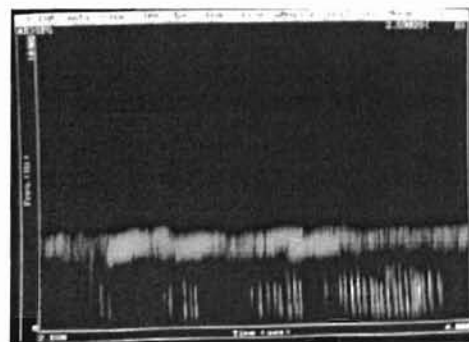
2.2.2 only processes those 256-point sample vectors. Compare to filtering and grouping algorithm that works on the whole data set (for 10 second that will be 80,000 point totally), that LPC method needs less computation. But this method depends on manually selecting candidate signals in the first stage, while the method proposed in this thesis is automatic. And the LPC method also requires further computation for pattern classification, which requires additional computation time.

As mentioned in Section 3.2, calls of type II (PSST) is the simple form of that of type I. So just by changing some parameters this filtering and grouping algorithm may also been used to identify type II frog calls.

Since many species of frogs tend to make sounds in a similar pitch/pulse repetition mode, it is convenient to use this filtering algorithm to identify these species. In doing this parameters such as passband of BPF and some criteria about pitch duration and call length need to be changed accordingly. Figure 3-7 shows some sample spectrograms of different species with similar call patterns.



Grey Treefrog



Northern Leopard Frog

Figure 3-7: Spectrograms of Two Typical Species [16]

There is no need to use the BPF and grouping algorithm to identify species with continuous calls like BUAM. Faster Fourier Transforms (FFT) is enough to accomplish this task. Main frequency band of BUAM calls is [1,600Hz 2,000Hz]. Usually peak frequencies of BUAM calls are within range [1,650Hz 1,850Hz]. Pitch repetition rate is roughly 0.04 ~ 0.05s, that is 320 ~ 400 points and the duration of one single pitch is usually 0.03 ~ 0.04s, that is 240 ~ 320 points. So a 512-point time interval at anytime within a call is sufficient to cover most parts of one single pitch. Specifically, a 512-point data set was extracted every 2 seconds along the data file. For one data file approximately 10 seconds long, totally four or five data sets can be extracted. FFT was performed on each data set. Then peak frequency values within [1,000Hz 2,000Hz] of each result were checked to see whether they are still within [1,650Hz 1,850Hz]. If so, there must be BUAM calling in this period.

Shown below is one example. For a given sound signal, totally four 512-point data sets were extracted and four different FFT have been calculated. The peak frequencies within [1,000Hz 2,000Hz] are 1,781Hz, 1,813Hz, 1,781Hz and 1,750Hz separately. All are within [1,650Hz 1,850Hz], which means there is BUAM calling in this period. The results of four FFT (here only first 257 points) are shown in Figure 3-8. In Figure 3-8, from left to right are results of FFT analysis of data segments extracted from original signal at 1s, 3s, 5s, 7s, respectively.

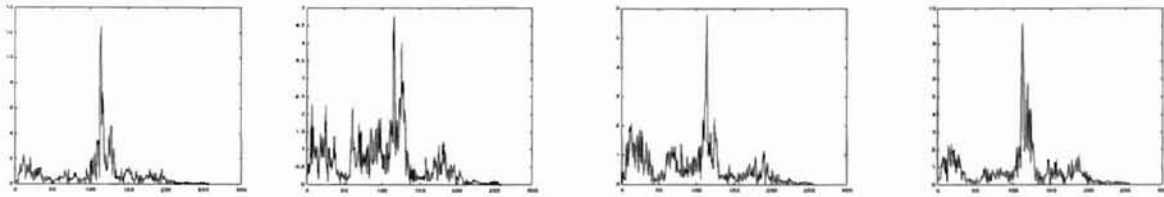


Figure 3-8: Results of Continuous Calculation of 512-point FFT Within One Data File

Ideally, this method is reliable at detecting continuous calls that have peak frequency values within a narrow frequency band. But if there is some kind of continuous noise, which happens to have peak frequency value in the same frequency range, a mismatch is unavoidable.



## CHAPTER IV

### IDENTIFICATION OF FROG INDIVIDUALS

#### 4.1 Individual Identification Overview

Here individual identification refers to identifying individual frogs within the same species and estimate the number of the identified frogs. Not all species of frogs are available for this task. The underlying assumption of automatic individual identification is that human experts may distinguish different individuals within the same species. If so, their knowledge can be used for analysis purposes and make the automatic identification possible. If human experts cannot tell the difference of one call from another, it is unlikely for machine to tell the difference. The reason is that in this case no prior knowledge is available. For example, one BUAM frog usually makes call continuously for several tens of seconds or even minutes. There may or may not be other individual calls at the same time. Experts cannot tell the difference. In this case without any specific sample or prior knowledge, no baseline can be established.

Within four species of interest only RAUT can be identified by human ear individually, which is proven by experts in zoology, so our research will focus on individual identification within this species only.

Before individual identification can be examined, typical samples of individual calls must be collected first. In the species identification stage, individual RAUT calls with several pitches have been extracted as seen in Figure 3-6. Since one pitch is the

basic element of each RAUT call, we may choose one typical pitch as the sample for one RAUT call. The duration of one single pitch ranges from 0.02s to 0.03s, that is 160 ~ 240-point data segment. A finite duration window  $w[n]$  is applied to original signal  $Y[n]$  prior to any signal analysis. This produces the finite length sequence  $v[n] = w[n]Y[n]$ . There are many kinds of windows in digital signal processing such as Bartlett, Hamming, Hanning, Blackman, Kaiser, and rectangular window. Here we choose one non-overlapping 512-point Hamming window. The center point of Hamming window was chosen to be at the maximum filtered value ( $Y_{bp}(t)$ ) within one call. Thus the strongest pitch within one call has been extracted and serves as the sample vector for this frog call. Figure 4-1 illustrates this process. The shaded area contains that strongest pitch.

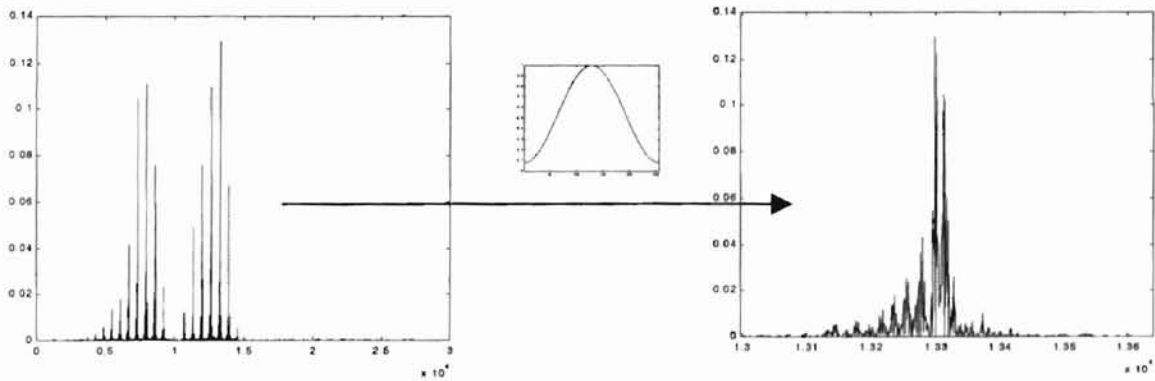


Figure 4-1: Filtered Signals (left) and Extracted Single Pitch (right)

For each RAUT call we can extract one 512-point data segment as its sample vector. All future works are based on analysis of this data segment.

## 4.2 Feature Extraction Algorithms

Feature extraction involves preliminary processing of signals to obtain suitable parameters that reveal distinguishable nature of the specific kind of signal. The aim of feature extraction is to devise a transformation that extracts the signal features hidden in the original domain. Corresponding to different characteristics of signals, different transformations should be properly selected to extract those most typical features from the original signal domain, thus to make the following step of signal analysis, which, in this case, the pattern classification, much easier. Three different algorithms used to extract typical features from those individual pitches will be discussed next. They are: time domain based method – Linear Predictive Coding (LPC); frequency domain based method – Time-Dependent Fourier Transform (TDFT) and time-scale domain based method – Wavelet Packet Transform (WPT).

### 4.2.1 Linear Predictive Coding Method

LPC is one of the most powerful speech analysis techniques. The theory of LPC, as applied to speech, has been well understood for many years [15]. The basic idea behind linear predictive analysis is that a new sample of the function under analysis can be estimated or predicted as a linear combination of a past number of samples of that function. By minimizing the sum of the squared error between the actual samples and the predicted ones over a finite interval of the function under test, a unique set of prediction

coefficients can be found. These coefficients used in the linear combination are the prediction coefficients. A linear predictor with prediction coefficients  $a_k$ , is defined as follows:

$$Z(n) = \sum_{k=1}^p a_k Z(n-k) + u(n) \quad (4.1)$$

where  $Z(n)$  is the  $n^{\text{th}}$  predicted sample,  $u(n)$  is the noise signal with zero mean and normal distribution, and  $p$  is the number of coefficients.

Speech is produced by excitation of an acoustic tube, the vocal tract, which is terminated on one end by the lips and on the other end by the glottis. There are three basic classes of speech sounds [26]:

- Voiced sounds – produced by exciting the vocal tract with quasi-periodic pulses of airflow caused by the opening and closing of the glottis.
- Fricative sounds – produced by forming a constriction somewhere in the vocal tract and forcing air through the constriction so that turbulence is created, thereby producing a noiselike excitation.
- Plosive sounds – produced by completely closing off the vocal tract, building up pressure behind the closure, and then abruptly releasing the pressure.

It is assumed that long-term non-stationary and time variant sound signals can be treated stationary and time-invariant over a short time interval on the order of 30 or 40ms. With a constant vocal tract shape, speech can be modeled as the response of a linear time-invariant system (the vocal tract) to a quasi-periodic pulse train for voiced sounds or wideband noise for unvoiced sounds. In voiced speech, the vocal tract, mouth and nose act as a filter that shapes the periodic excitation from the vocal cords [11]; this is the

source-filter model of speech production. LPC provides a good model of the speech signal. This is especially true for the quasisteady state voiced regions of speech in which the all-pole model of LPC provides a good approximation to the vocal tract spectral envelope [33]. The mechanism that frogs use to make advertisement calls are similar to human beings [6]. Also the length of 512-point sample is 64ms, similar to those voiced signals used for linear predictive analysis. Thus in our case we can use LPC for frog call analysis purpose.

There are different ways of computing LPC coefficients, such as the covariance method, the autocorrelation formulation, the lattice method, the inverse filter formulation, the least square method, the spectral estimation formulation, the maximum likelihood formulation, the inner product formulation, and the neural network method [33]. Here we use the least square (LS) method which is widely used in estimation theory [21].

To use the LS method,  $p$ , the number of LPC coefficient, should be determined first. Generally  $p$  should not be too large while ensuring low mean square error of the predicted and actual signal value. Also since later we need to calculate FFT of the LPC coefficients,  $p$  is usually chosen to be a power of 2.

After  $p$  has been determined, LS estimate first determines  $N+1$  time domain LPC filter coefficients by LS estimation:

$$\begin{bmatrix} Z(p+1) \\ Z(p+2) \\ \vdots \\ Z(N) \end{bmatrix} = \begin{bmatrix} Z(p) & \cdots & Z(2) & Z(1) \\ Z(p+1) & \ddots & & Z(2) \\ \vdots & & \ddots & \vdots \\ Z(N-1) & \cdots & \cdots & Z(N-p) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} + \begin{bmatrix} u(p+1) \\ u(p+2) \\ \vdots \\ u(N) \end{bmatrix} \quad (4.2)$$

Equation (4.2) corresponding to:

$$Z = Ha + V \quad (4.3)$$

where

$$Z = \begin{bmatrix} Z(p+1) \\ Z(p+2) \\ \vdots \\ Z(N) \end{bmatrix} \quad H = \begin{bmatrix} Z(p) & \cdots & Z(2) & Z(1) \\ Z(p+1) & \ddots & & Z(2) \\ \vdots & & \ddots & \vdots \\ Z(N-1) & \cdots & \cdots & Z(N-p) \end{bmatrix} \quad a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} \quad V = \begin{bmatrix} u(p+1) \\ u(p+2) \\ \vdots \\ u(N) \end{bmatrix} \quad (4.4)$$

The solution based on LS estimate is:

$$\hat{a}_{LS} = [H^T H]^{-1} H^T Z \quad (4.5)$$

After the LPC coefficients have been obtained, we need to calculate the FFT of the  $p$ -point LPC coefficients and get  $(p/2) + 1$  unique spectral magnitudes as final feature vectors. When  $p$  equals to 16, the dimension of final feature vectors of one frog call will be 9.

#### **4.2.2 Time-Dependent Fourier Transform Method**

Perhaps the most well-known signal analysis tool is Fourier analysis, which breaks down a signal into constituent sinusoids of different frequencies. In other words, it transforms signal from time domain to frequency domain. Basic Fourier based methods like DFT (discrete Fourier transform) is widely used in analyzing the frequency content of continuous-time sinusoidal signals. But since time information is lost during transformation, it may cause serious problems when dealing with nonstationary signals.

The need for multiplication of signal  $x[n]$  by one window  $w[n]$  is a consequence of the finite-length requirement of the DFT. For nonstationary signals such as speech

signals, the signal properties (amplitudes, frequencies, and phased) will vary with time. A single DFT estimate is not sufficient to describe such signals. As a result, the time-dependent Fourier transform (TDFT), also referred to as the short time Fourier transform (STFT), is usually used to analyze this kind of signals.

The TDFT of a signal  $x[n]$  is defined as

$$X[n, \lambda] = \sum_{m=-\infty}^{\infty} x[n+m]w[m] e^{-j\lambda m} \quad (4.6)$$

where  $w[n]$  is a window sequence. In the TDFT representation, the one-dimensional sequence  $x[n]$  is converted into a two-dimensional function of the time variable  $n$ , which is discrete, and the frequency variable  $\lambda$  is continuous.

In this thesis, a fixed window length was used in Fourier analysis. Calculate  $N$  point FFT of the windowed data set and we can get totally  $N/2 + 1$  point unique frequency vectors.

The primary purpose of the windowing in TDFT is to limit the extent of the sequence to be transformed so that the spectral characteristics are reasonably stationary over the duration of the window. The more rapidly the signal characteristics change, the shorter the window should be. A long analysis window is not suitable in analyzing short duration bursts, while a short analysis window is not appropriate for long duration component. Thus the choice of proper window length experiences typical trade-offs between the frequency resolution and the time resolution quality. Once you choose a particular size for the time window, that window is the same for all frequencies. This deficiency of TDFT naturally leads to the Wavelet analysis, which is more efficient than Fourier based analysis for nonstationary signals.

### 4.2.3 Wavelet Packet Transform Method

Many signals require a more flexible approach – one where we can vary the window size to determine features of these signals more accurately either in time or frequency. Obviously TDFT cannot satisfy this requirement. The Wavelet analysis represents the next logical step: a windowing technique with variable-sized regions. Wavelet analysis allows the use of long time intervals where we want more precise low frequency information and short regions where we want high frequency information.

Figure 4-2 shows what this looks like in contrast with the time-based, frequency-based, and TDFT views of a signal [24]:

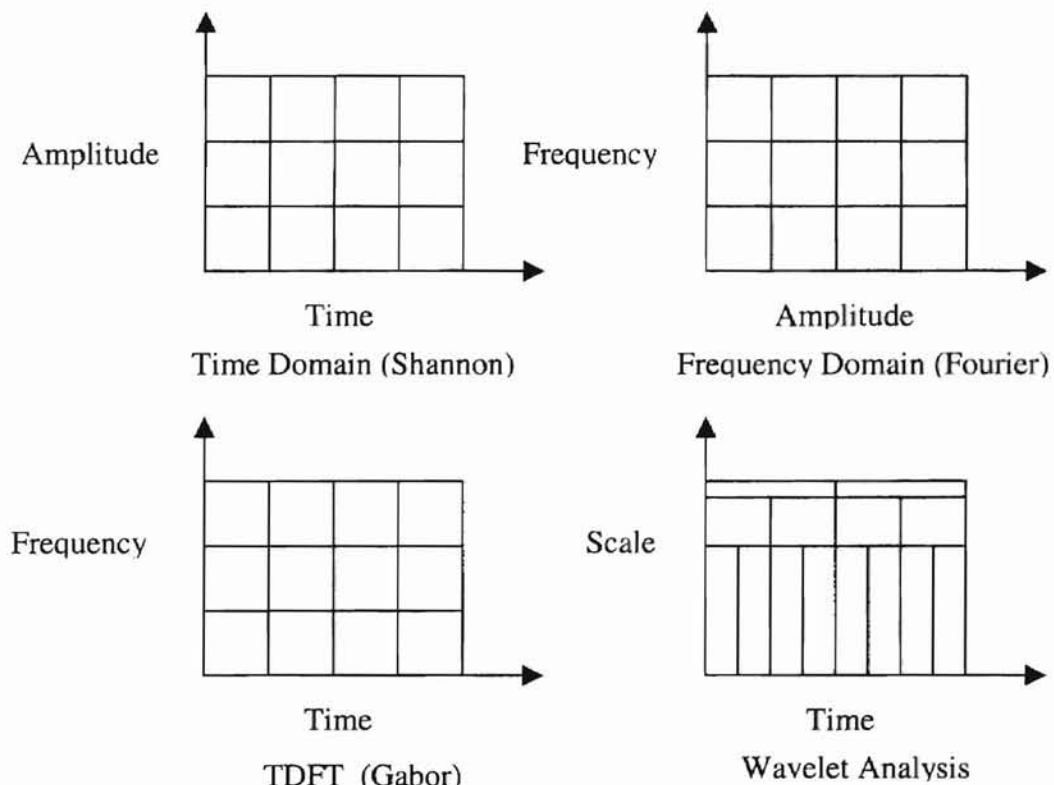


Figure 4-2: Wavelet Analysis Versus Time Domain, Frequency Domain and TDFT

Analysis



Wavelet analysis does not use a time-frequency region, but rather a time-scale region. It is capable of revealing aspects of data that other signal analysis techniques miss, aspects such as trends, breakdown points, discontinuities in higher derivatives, and self-similarity [24]. Though wavelets have a brief history, they have already proven themselves to be an indispensable addition to the analyst's collection of tools. Wavelets provide better time-frequency resolution. They are more efficient than Fourier based methods for non-stationary signal analysis.

A wavelet is a waveform of effectively limited duration that has an average value of zero. Different from Sinusoids, wavelets are irregular and asymmetric. Wavelet analysis breaks up a signal into shifted and scaled versions of the original wavelet (mother wavelet). One typical wavelet, the 8<sup>th</sup> Daubechies (db8) is shown in Figure 4-3.



Figure 4-3: db8 Wavelet

The Continuous Wavelet Transform (CWT) is defined as the sum over all time of the signal  $f(t)$  multiplied by scaled, shifted versions of the wavelet function  $\Psi$ :

$$C(\text{scale}, \text{position}) = \int_{-\infty}^{\infty} f(t)\Psi(\text{scale}, \text{position}, t)dt \quad (4.7)$$

The results of the CWT are many wavelet coefficients  $C$ , which are a function of scale and position. Multiplying each coefficient by the appropriately scaled and shifted wavelet yields the constituent wavelets of the original signal.

If we choose scales and positions based on powers of two – so called dyadic scales and positions – then our analysis will be expected to be more efficient and accurate. An efficient way to implement this scheme using filters was developed in 1988 by Mallat [14]. The Mallat algorithm is known in the signal processing community as a two-channel subband coder [37]. This very practical filtering algorithm yields a fast wavelet transform (FWT). The filtering process, at its most basic level, looks like Figure 4-4.

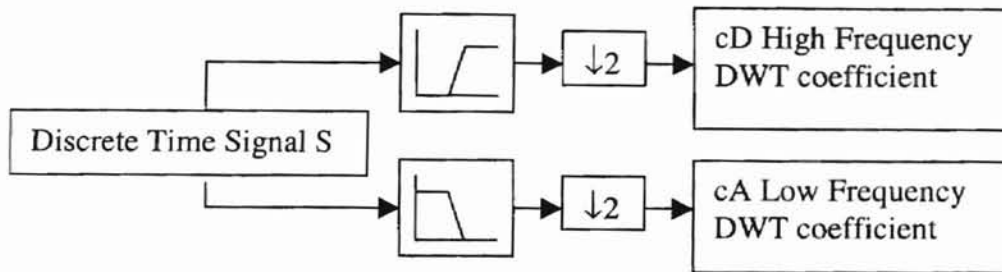


Figure 4-4: Illustration of DWT of Signals

The original signal first passes through two complementary filters (one high pass filter and one low pass filter), then after downsampling (downsampled by 2), produces DWT coefficients. Here cD means detail part from high frequency content of signal and cA means approximation part from low frequency content of signal.

Wavelet decomposition process can be iterated so that one signal is broken down into many lower resolution components. This is called the wavelet decomposition tree. In

level 1 the original signal is decomposed into an approximation and detail. The approximation itself is then split into a second level approximation and detail, and the process is repeated.

The Wavelet Packet Transform (WPT) is a generalization of wavelet decomposition that offers a richer range of possibilities for signal analysis. Not only the details by the approximations can be split as well. Figure 4-5 shows one wavelet packet decomposition (WPD) tree.

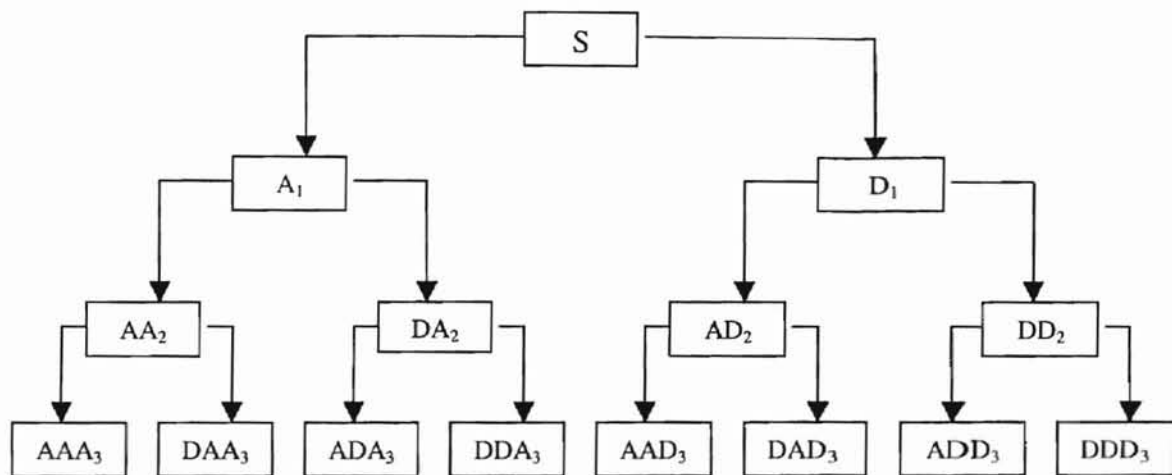


Figure 4-5: Three Level Wavelet Packet Decomposition Tree

Whereas the DWT only decomposes low frequency components of signal, WPT decomposes both the low frequency and high frequency components. This rich abundant information with arbitrary time frequency resolution can allow extraction of features that combine both stationary and nonstationary characteristic. However, one deficiency that wavelet bases inherently possess is the lack of translation invariant property. A signal with a time shift does not result in the time shifted wavelet packet coefficient. Direct

assessment from all wavelet packet coefficients often turns out to be tedious or leads to inaccurate results. Here we introduce the idea of wavelet packet node energy [40]. Image representation as shown in Figure 4-6 was used to represent the WPD tree shown in Figure 4-5 for interpretation purpose. Each cell  $w(i, j)$  refers to one node in the decomposition tree, here  $i$  is the scaling parameter and  $j$  is the oscillation parameter. We call each  $(i, j)$  as a wavelet packet node. Each  $(i, j)$  has  $K$  coefficients,  $w_{i, j, k}$ . For one signal of length  $2^N$ , the maximum value of  $i$  is  $N$ , for each  $i$ , the maximum value of  $j$  is  $2^i - 1$  and the maximum value of  $k$  is  $2^{N-i}$ .

w(0,0)							
w(1,0)				w(1,1)			
w(2,0)		w(2,1)		w(2,2)		w(2,3)	
w(3,0)	w(3,1)	w(3,2)	w(3,3)	w(3,4)	w(3,5)	w(3,6)	w(3,7)

Figure 4-6: Node Representation of WPD Tree

The wavelet packet node energy is defined as:

$$e_{i,j} = \sum_k W_{i,j,k}^2 \quad (4.8)$$

which measures the signal energy contained in some specific frequency band indexed by parameters  $i$  and  $j$ . In our case each wavelet packet node energy value was defined as an individual feature component and was used as a robust rudimentary exploration of the specific signal features. For an  $r$  level decomposition, we can get totally  $2^1 + 2^2 + \dots + 2^r = 2^{r+1} - 2$  sets of node energy coefficients. These coefficients are final results extracted from each frog call signal by using wavelet method.

### 4.3 Dimension Deduction/Feature Selection Algorithm

In Section 4.2 three feature extraction algorithms were introduced. It can be noticed that using TDFT and WPT may produce high dimension feature vectors for each individual frog call signal. For example, if a window of length 512 was used, then one TDFT feature vectors will contain totally 257 coefficients and the dimension of one WPT feature vectors with full decomposition will reach 510. Direct manipulation on whole data set is not feasible because of high dimensionality of data and the existence of undesired components that make the classification unnecessarily difficult. Thus it is desirable to use lower dimensional feature vectors as input for the pattern classifier.

To reduce dimension of feature vectors, one idea is to find a linear transformation that maps high dimensional data onto lower dimensional space, the other is to select those feature components that contain most discriminant information and discard those provide little information which is useful for classification purpose [10]. Here the second method is chosen for dimension reduction purpose.

Specifically, the feature component  $\{f_k | k=1, 2 \dots n\}$  is ranked:

$$J(f_1) \geq J(f_2) \geq \dots \geq J(f_d) \geq \dots \geq J(f_n) \quad (4.9)$$

where  $J(\cdot)$  is a criterion function for measuring the discriminant power of a specific feature component,  $f_k$ . Feature subset can be chosen from those features having larger criterion function values.

The concept of probabilistic structure of classes was introduced to begin the discussion. Given two probability density function (PDF) of class  $c_1$  and  $c_2$ , for one specific feature variable  $x$ , if  $p(x|c_1)$  is zero for all  $x$  such that  $p(x|c_2)$  is not zero as shown

in Figure 4-7(a), then we say these two classes are fully separable. Or if  $p(x|c1)$  equals to  $p(x|c2)$  everywhere, we say these two classes are not separable. Intuitively, a criterion function could be set as measurement of overlapping between  $p(x|c1)$  and  $p(x|c2)$ . More overlapping of these two functions means lower value of this criterion function (discriminant power).

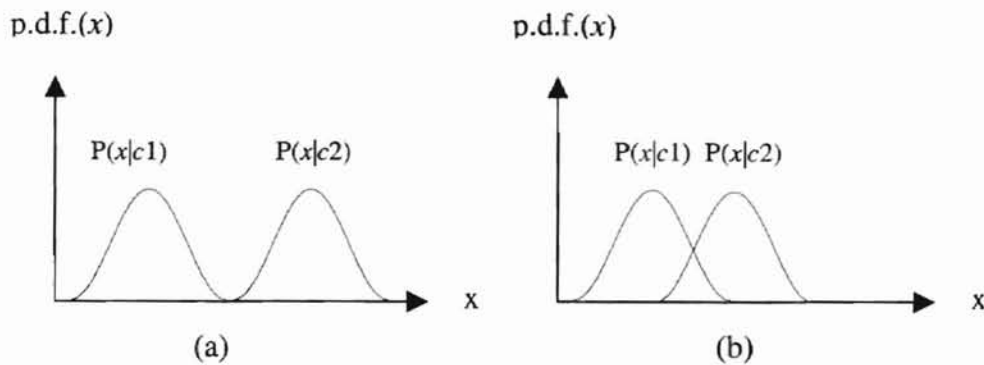


Figure 4-7: PDF of (a) Two Well-separated Classes and (b) Two Overlapping Classes

Generally speaking, a criterion function that measures the overlap between two classes has the following properties [10]:

- The measure is minimum when the conditional PDF for class  $c1$  and  $c2$  are identical, i.e.  $J(\cdot) = 0$ , if  $p(x|c1) = p(x|c2)$ .
- The measure is non-negative.  $J(\cdot) \geq 0$ .
- The measure attains a maximum when the classes are disjoint, i.e.  $J(\cdot) = \max$ , if  $p(x|c1) = 0$  wherever  $p(x|c2) \neq 0, \forall x$ .

Properties above only give an intuitive justification of their suitability for feature selection. Some criteria that provide a direct indication of the amount of the overlap of the class probability densities are listed below [10]:

Chernoff distance:

$$J(.) = -\ln \int p^s(x|c1) \cdot p^{1-s}(x|c2) dx, s \in [0,1] \quad (4.10)$$

where  $s$  is a parameter between 0 and 1.

Matusita distance:

$$J(.) = \left\{ \int [p(x|c1)^{1/2} - p(x|c2)^{1/2}]^2 dx \right\}^{1/2} \quad (4.11)$$

In this study, another simple while efficient criterion function known as Fisher's criterion [41] was adopted. For a two classes problem it is given by:

$$J_{f_k}(i, j) = \frac{|\mu_{i,f_k} - \mu_{j,f_k}|^2}{\delta_{i,f_k}^2 + \delta_{j,f_k}^2} \quad (4.12)$$

where  $\mu_{i,f_k}$  and  $\mu_{j,f_k}$  are the mean values of  $k^{\text{th}}$  feature,  $f_k$ , for class  $i$  and  $j$ ;  $\delta_{i,f_k}^2$  and  $\delta_{j,f_k}^2$  are the variance of the  $k^{\text{th}}$  feature,  $f_k$ , for class  $i$  and  $j$  correspondingly. For multiple class (class number equals to  $L$ ) case, the general approach is to take summation of the pairwise combinations of  $J_{f_k}(i, j)$ :

$$J_{f_k} = \sum_{i=1}^{L-1} \sum_{j=i+1}^L J_{f_k}(i, j) \quad (4.13)$$

as an estimation of discriminant power for the specific feature  $f_k$ . Eq. (4.12) provides a measure to evaluate the effectiveness of the "global" feature that is simultaneously suitable to differentiate all classes of signals. For small classes case, this approach may be sufficient. When the number of classes increases, this equation becomes more ambiguous. A large value of  $J_{f_k}$  may be due to the accumulations of many relatively small values (an unfavorable case) or to a few significant terms with negligible majority (a favorable case). Also a feature with large  $J_{f_k}(i, j)$  value to class  $i$  and  $j$  may have very

small discrimination power for other classes, thus makes  $J_{f_k}$  also very small. To avoid these problems, two methods are established as possible alternatives.

Method I:

Instead of trying to select features, which are effective for the entire multi-class problem globally as measured by Equation (4.12), a feature subset based on Equation (4.11) was selected for each possible pair of classes [40]. Then the union of feature components selected from each pair of classes was taken to form the final feature vector. Specifically, given a  $L$ -class problem with  $n$  feature components, the process is detailed in the following steps:

1. For each possible class pair  $\{(i, j) \mid i = 1, 2, \dots, L-1, j = i+1, i+2, \dots, L\}$ , calculate the discriminant power measure for each feature component,  $f_k$ , using Equation (4.12).
2. For each class pair, sort  $J_{f_k}(i, j)$  such that:

$$J_{f_1}(i, j) \geq J_{f_2}(i, j) \geq \dots \geq J_{f_d}(i, j) \geq \dots \geq J_{f_n}(i, j) \quad (4.14)$$

Determine the feature subset  $F_{i,j}$  for each class pair by selecting  $d$  feature components that have maximum  $J_{f_k}(i, j)$  value:

$$F_{i,j} = \{f_k \mid k = 1, 2, \dots, d\}, i = 1, 2, \dots, L-1; j = i+1, i+2, \dots, L. \quad (4.15)$$

3. Form the final feature set by taking the union of each feature subset.

$$F_{final} = \left\{ \bigcup_{i=1}^{L-1} \bigcup_{j=i+1}^L F_{i,j} \right\} \quad (4.16)$$



Method II:

This method is based on similar idea of method I. Compare to method I, method II may choose different number of feature components from each class pair, thus more reasonable feature components are expected with this method.

The first step is the same as that of method I. In the second step, after  $J_{f_k}(i, j)$  were sorted in descending order, the whole data set was normalized. That is:

$$J = \sum_{k=1}^n J_{f_k}(i, j) \quad (4.17)$$

$$J'_{f_k}(i, j) = J_{f_k}(i, j) / J; k = 1, 2, \dots, n \quad (4.18)$$

Set one threshold value  $H \in (0, 1]$ . Determine the feature subset  $F_{i,j}$  for each class pair by selecting  $D$  feature components that have maximum  $J_{f_k}(i, j)$  value.  $D$  must satisfy:

$$\begin{cases} \sum_{k=1}^D J'_{f_k}(i, j) \geq H \\ \sum_{k=1}^{D-1} J'_{f_k}(i, j) < H \end{cases} \quad (4.19)$$

$$F_{i,j} = \{f_k \mid k = 1, 2, \dots, D\}, i = 1, 2, \dots, L-1; j = i+1, i+2, \dots, L. \quad (4.20)$$

The third step is the same as that of method I. By using method II, different number of feature components may be extracted from each class pair. If two classes are well separable, there must exist a few feature components that contain much larger  $J_{f_k}(i, j)$  value than most other feature components. If two classes are not well separable, then most feature components tend to have similar relatively small  $J_{f_k}(i, j)$  values. By setting this threshold, we may choose fewer feature components from well separable class pairs and more feature components from those classes that were difficult to separate. Thus we

may get feature subsets with relatively low dimension while still contain high discriminant power.

#### **4.4 Pattern Classification Algorithm**

Once after suitable feature components have been extracted from original feature set, it is then necessary to determine individual frogs based upon these features. Artificial neural networks (ANN) and a variety of multivariate statistical methods have been used for pattern recognition/classification problems similar to that of this research [2]. In this study, a static network (Multilayer Perceptron, MLP) was used for classification purpose.

Neural network classifiers are widely used in pattern recognition problems because they are universal function approximators and because of their nonlinear nature, they have the ability to capture the underlying non-linearity from the incoming data.

However, for MLP, existing pattern must be used to train the network and the classifier can only detect those already existing classes. That is, in our case, we already know how many individuals of RAUT in one area and our classifier can only detect calls of these identified frogs. If one new RAUT frog makes a call and its call features are fed into the classifier, the MLP classifier may not identify it as a new one and possibly will classify it into one already existing RAUT frog class.

To overcome this problem, one method was introduced here. The fuzzy neural network based classifier is called Incremental Learning Fuzzy Neuron Network (ILFN) [20]. It uses incremental learning algorithm and can detect new classes of patterns and

update its parameters while in an operating mode. It has an on-line (real-time) and fast learning algorithm without knowing a priori information. In addition, it has the capability to make soft (fuzzy) and hard (crisp) decisions, and it is able to classify both linear separable and nonlinear separable problems. By using ILFN, new individual frogs can be detected in real time and there is no need to re-train the network with known features.

## CHAPTER V

### TEST RESULTS

In this chapter, natural frog calls collected via Telinga Pro V Mono Parabolic microphone mounted at various lakesides in Stillwater, Oklahoma were used to validate the feasibility of the proposed species and individual identification methods. SONY PCM-M1 digital audio recorder was used to store original audio signal into digital audiotape (DAT). The Voyetra Turtle Beach system, including Montego II sound card and Turtle Beach AudioStation 32 software, was used to transfer audio signals stored in DAT into personal computer (PC) with a digitized WAVE format. All test programs were written to run under MATLAB version 5.3 or higher. A Pentium III 500 PC hosted all the programs.

#### 5.1 Results for Species Identification

For species identification, one DAT with total length of 50 minutes was chosen as sample. It contains frog calls of all four species obtained from several lakesides within the State of Oklahoma. Each species contains several different individuals. The entire DAT data were manually saved into PC with WAVE format. Each file segment was approximately 10 seconds long. Each data set was fed into four different programs that identify one species correspondingly. The goal is that each program may be able to

identify all clear calls of that species and does not count other signals (e.g., calls made by other species).

It is much more desirable for our system to fail to recognize a call (a false negative) than to incorrectly indicate the call of a particular species is present (a false positive). It is crucial then to choose parameters such that false positives are minimal [39]. For example, to identify RAUT and PSCL frogs, according to call properties generalized in Chapter III, we narrow down the ranges of pitch duration thus to avoid mismatch with other short duration spikes. Also, in clustering algorithm, to choose possible call signals and discard the false impulses we do thresholding on the squared signal  $Y_{sqr}(t)$ . If the threshold value is too big, more irrelevant signals will be discarded but some portion of true frog call signals (those pitches in the beginning or in the end of one call) will also be thrown away due to little energy they have. There exists a trade-off between recognizing more false negative and less false positive.

In practice, by carefully adjusting various parameter values, the result for species identification is quite promising. Within this sample period there are hundreds of calls belonging to four different species. Except for some weak calls and some calls obscured by environmental noise, most clear calls can be detected and identified as belonging to correct class with nearly 100% accuracy. For frog species of BUAM and PSST, the results are found to be perfect.

Yet, there do exist a few mismatches when identifying species RAUT and PSCL. The noise causes part of the problem. In most natural situations background noise is extremely high and its temporal and spectral structure are complex and variable. In this DAT tape, there always exist three types of noise:

1. Noises made by other living creatures: Including calls made by other frog species and some insects like crickets. Occasionally there exist some dog barks and human speech if the pond is close to human community.
2. Noises made by natural phenomena: Including wind noise and rain noise.
3. Noises made by vehicles: Including noises made by automobiles.

Among these three types of noises, 1 and 2 occur more frequently. If the frequency band of these noises is different from that of the specific frog species, they can be removed in the stage of filtering. Or if the spectrogram of these noises didn't appear to be a steady pulse repetition mode just like those of RAUT and PSCL, they can also be eliminated in the stage of clustering. But if the main frequency band and pulse shape of that kind of noise are quite similar to those of frog species, a mismatch is inevitable by using the proposed identification method. That is, the occurrence of one that kind of noise signal may be mistakenly determined to be one certain kind of frog call.

For the third type of noise, although it happens occasionally, if the noise level is high, frog calls may be occluded and sometimes mismatch may also occur.

## **5.2 Results for Individual Identification**

Here individual identification was focused on species RAUT, because this was the only species human experts can distinguish different individuals according to their sounds. Only with this prior knowledge, the proposed individual identification method can function properly.

### **5.2.1 Data Segmentation**

Individual identification is based on the results of species identification. After one call of RAUT species has been identified, its calling period has also been determined simultaneously. Then a non-overlapping 512-point Hamming window was used to extract 512-point time series data segment from one RAUT call as its sample vector. The length of window guarantees to contain at least one pitch (the strongest one) within this sample vector. For the same species, it is reasonable to assume that call patterns of different individuals can be fully explored by analyzing one single pitch.

Before this data segment can be used for further analysis, the mean value of this segment is calculated first and subtracted from the whole data set. Because signals with non-zero mean may produce incorrect spectrum estimate especially in low frequency band, subtracting mean value from the signal often leads to a better estimate at neighboring frequencies [26].

### **5.2.2 Generation of Training / Testing Data Set**

Because frogs are sensitive to sudden changes of environment, their calls are difficult to collect. Also human experts usually have limited capability to identify different individuals. For these reasons there are totally 66 data sets been identified, which correspond to four different individual RAUT frogs. For these 66 data sets each time 44 data sets were randomly chosen as training data while the remaining 22 data sets

were used in testing. The distribution of training and testing data sets are shown in Table 5-1.

Table 5-1: Number of Individual Samples and the Distribution of Training / Testing Sets

RAUT Individual	Total number of data sets	# of data sets for training	# of data sets for testing
Frog 1	16	11	5
Frog 2	20	13	7
Frog 3	17	11	6
Frog 4	13	9	4

### **5.2.3 System Description**

After 66 numbers of 512-point feature vectors are extracted, Linear Predictive Coding (LPC), Time-Dependent Fourier Transform (TDFT), and Wavelet Packet Transform (WPT) are used for feature extraction. For TDFT and WPT, two different dimension reduction algorithms are used to derive the final feature vector, which will be fed into a neural network classifier. The steps provided for each method are summarized below.

LPC:

1. Determine number of LPC coefficient,  $p$ , cording to mean square error (MSE).



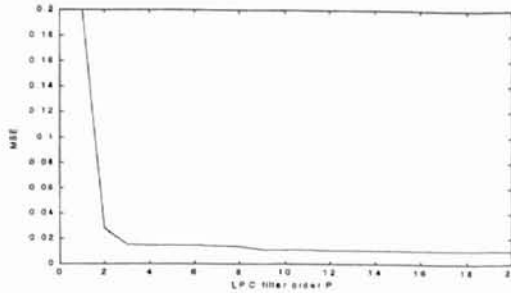


Figure 5.1: The Relationship Between LPC Coefficient Number  $p$  and MSE

If we determine  $p$  first and use Equation (4.5) to compute the LPC coefficients, we can determine the filtered output of LPC filter. Then compute the mean square error, which indicates the difference between filtered value and actual value. In this way we establish the relationship between LPC filter of order  $p$  and the corresponding MSE. Figure 5.1 is based upon the calculation made on one data set. In fact all data sets show similar relationship between  $p$  and MSE. Based on this figure,  $p$  was chosen to be 16. In this design  $p$  is not too big but ensures low MSE.

2. For each sample vector, determine 16 time domain LPC filter coefficients by using Equation (4.5).
3. Calculate FFT of 16-point LPC coefficients and obtain 9 unique spectral magnitudes.
4. Normalize these 9 spectral magnitudes to get final feature vector.

TDFT:

1. Calculate 512-point FFT for each windowed data set and obtain 257-point spectral magnitude vector.

2. Use feature extraction method I and II, set parameters  $d$  (defined in method I) and  $H$  (defined in method II) and derive the corresponding feature subsets (expect to contain feature components with most discriminant power).
3. Normalize these two feature subsets and acquire the final feature vectors.

WPT:

1. Perform eight-level wavelet packet decomposition for each 512-point data set by using Daubechies 8-point wavelet function.
2. Calculate wavelet packet node energy according to Equation (4.8) and get one 510-point feature vector.
3. Use feature extraction method I and II to derive the corresponding feature subsets.
4. Normalize these two feature subsets and obtain the final feature vectors.

In the last step of TDFT and WPT methods, feature vectors were normalized before they are fed into a neural network classifier. The reason of normalization is to maintain the similar distances between feature vectors. The normalization was achieved by the following operation,

$$\hat{x} = \frac{x - \mu}{\delta}, \quad (5.1)$$

where  $\hat{x}$  is the normalized version of vector  $x$ ,  $\mu$  is the mean value and  $\delta^2$  is the corresponding variance.  $\mu$  and  $\delta^2$  are estimated from the feature subset. In this way, the resulting vector set will have zero mean and unity variance.

### **5.2.4 Test Results**

After training data were obtained by three feature extraction algorithms, LPC, TDFT and WPT methods, as well as two dimension reduction/feature selection algorithms, method I and method II, they were fed into a neural network classifier. By checking the final results of the classifier to those testing data sets, some conclusions may be drawn pertaining to which method is better and which one is not.

For dimension reduction methods I and II, parameter  $d$  and  $H$  should be carefully chosen so that feature vectors with same or similar dimension may be generated.

There are totally 66 data sets available. In each test 44 of them were randomly chosen as training sets while the remaining 22 were used as testing sets. This process is repeated 1,000 times. The number of training sets and testing sets within one class (calls produced by the same frog) are fixed, as seen in Table 5.1. The mean value of these 1,000 simulations was calculated as indicator for test performance. The variance is also computed. If all variances are not high and in the same level, then the performance of the whole system can be regarded as stable.

The neural network used here is MLP. The network architectures are N-N-4 (with N neurons in the only hidden layer) and N-10-10-4 (with 10 neurons in the first hidden layer and 10 neurons in the second hidden layer), where N is the dimension of the final feature vector. In the learning phase, the network is trained until the mean square error is below 0.001, or the maximum epochs (set to 1,000) is reached. The resilient backpropagation algorithm (RPROP) [34] is used to train the network. In training we can make the desired output of MLP to be a perfect decision, i.e. one 1 and three 0s. But the

classifier will not produce such perfect decision in testing process. Usually each output will be between 0 and 1. Here we use the maximum output value as the most likely individual frog. In all cases a clear winner can always be identified. The classification results are shown in Table 5.2 - 5.4. Mean is referred to mean accuracy. It has range from 0 to 1. Var. is referred to as variance of the 1,000 runs. The training accuracy is always 100% in all cases and the corresponding variance is 0. So these tables only show test results for different methods.

Table 5-2: Test Results for LPC Method

LPC N = 9		N-N-4
	Mean	0.5068
	Var.	0.0104

Table 5-3: Test Results for TDFT Method

TDFT	N		N-N-4	N-10-10-4
Method I d = 4	17	Mean	0.6082	0.6089
		Var.	0.0075	0.0099
Method II H = 0.1	18	Mean	0.6218	0.6188
		Var.	0.0115	0.0089
Method I d = 8	33	Mean	0.6471	0.6505
		Var.	0.0093	0.0098
Method II H=0.16	33	Mean	0.6330	0.6377
		Var.	0.0094	0.0093

Table 5-4: Test Results for WPT Method

WPT	N		N-N-4	N-10-10-4
Method I d = 5	19	Mean	0.6827	0.6809
		Var.	0.0097	0.0105
Method II H = 0.06	18	Mean	0.7118	0.7164
		Var.	0.0068	0.0100
Method I d = 8	31	Mean	0.6609	0.6891
		Var.	0.0095	0.0102
Method II H=0.1	33	Mean	0.7218	0.6955
		Var.	0.0076	0.0150

First examine the results of LPC method. The classifier can only correctly classify roughly half of the test samples, which is not good and much lower than the results of TDFT and WPT methods. In Chapter II, some papers using LPC method to perform *species identification* were reviewed. Often they can reach 70% — 80% accuracy, which is much higher than what was realized here for *individual identification*. Also it can be noticed that these samples contain a large amount of noise. A rough estimate to some data files shows an average SNR (signal to noise ratio) of  $-3\text{dB}$ . The noise significantly deteriorates the performance of LPC filter and finally leads to poor performance of the neural network classifier.

To compare the performance of TDFT and WPT method, the following conditions have been set.

1. The dimensions of final feature vectors should be the same (i.e. 18 Vs. 18) or quite similar (i.e. 31 Vs. 33).
2. The feature selection methods should be the same (i.e. method I Vs. method I).

3. The network structures should be the same (i.e. N-N-4 Vs. N-N-4).
4. The variances should be similar and not too high (i.e. 0.0076 Vs. 0.0094).

Based on these conditions, the performance of TDFT and WPT methods was compared one by one. On the average, neural network classifiers based on WPT method acquires the accuracy of classification 8% higher than those based on TDFT method. This shows the conclusion mentioned in Chapter IV that Wavelet based method (in this thesis, identified as WPT) provide a better time-frequency resolution and they are more efficient than Fourier based methods for non-stationary signal analysis (in this case, frog call signals are surely non-stationary signals).

To compare the performance of these two feature selection methods I and II, similar to the way TDFT and WPT methods are compared, some conditions have been set. Only these two methods with the same (or similar) conditions are compared. Basically, when the dimension of final feature vectors are the same or quite similar, using method II may extract feature components with more discriminant power thus to make the performance of neural network classifier better. This is especially true in WPT case, in which by using method II the accuracy of classifier is on the average 3% higher than that of using method I. This also substantiates our assumption that we may choose fewer feature components to distinguish those easily separable classes and choose more feature components to distinguish those relatively not so easily separable classes. By this way more feature components that contains most discriminant power may be included with limited feature vector dimension.

It is observed that WPT method exerts a large amount of computation load compared to TDFT method. If WPT method is to be used, it is preferred to use low dimensional feature vector and use simple neural network structure. Among all these combinations one good solution can be found. That is, use WPT and feature selection method II (set  $H = 0.06$ ), get 18-point feature vectors, then use 18-18-4 MLP as classifier. Thus a considerable amount of computation is avoided, keeping the accuracy for classification remains high.

## CHAPTER VI

### CONCLUSIONS AND FUTURE WORK

#### 6.1 Conclusions of the Research

This thesis has investigated the feasibility of building an automatic frog call monitoring system based on in-field acquisition of sound signals. The frog species identification has been realized in the first stage. Different algorithms including filtering and grouping are developed to identify different species. The individual frog identification of species RAUT has been performed in the second stage. Since most of the research in the field of animal sound recognition are focused on species identification, the individual identification approach proposed in this thesis is novel. Three feature extraction algorithms including LPC, TDFT, and WPT, two dimensionality reduction algorithms (method I and II), and the neural network (MLP) classifier have been synergistically integrated together to facilitate the estimation of the population within the species of interest.

In Chapter II, some of the existing literature in the field of animal sound recognition were reviewed. Most are focused on species identification. Three digital signal pre-processing approaches and three pattern classification approaches are discussed in this chapter.

In Chapter III, the proposed species identification algorithm was discussed. For four different classes of frogs in the State of Oklahoma, three different methods have



been developed. For species with the most sophisticated call patterns, the bandpass filter was used first to get rid of most irrelevant noises and to lower the noise level. Then, according to the characteristics found in the spectrogram of the proposed frog call, grouping algorithm was employed to isolate periods of single pitches first and then to group the time interval of that frog call. By this way not only the desired class can be identified but its call interval can also be acquired. This method requires less computation compared to other approaches reported in the literature.

In Chapter IV, individual identification of RAUT frog was implemented step by step. After windowing the entire calling interval of one call acquired in species identification period, a 512-point data set was extracted as a sample vector. Three different feature extraction methods, LPC, TDFT, and WPT were used to extract feature vectors first. Since the dimensions of feature vectors derived from TDFT and WPT approaches are huge, two different dimensionality reduction/feature selection algorithms were proposed to obtain the feature subsets, which contains feature components with most discriminant power. Finally MLP, the neural network classifier, was used for classification and identification purposes.

In Chapter V, species identification algorithms discussed in Chapter III were used to test one 50 minute-long data set collected in field. Except for a few expected mismatches, the results were found to be nearly perfect. Then, combinations of different methods illustrated in Chapter IV were used to totally test 66 sample frog calls. By comparison, we found WPT was the best feature extraction method and the performance of dimensionality reduction method II was far better than method I.

## 6.2 Suggestions for Future Work

Although simulation results for species identification were reasonably satisfactory in this study, there still exist a few mismatches. Avoidance of these mismatches is not a trivial problem. Simply tuning the values of some parameters in the algorithm is not enough. A combination of a new algorithm with those already exist such as knowledge-based system is one possible solution. Considering the requirement for speed, the method used should be carefully selected.

Due to limited samples available in individual identification test examples, though we employed statistical measures to remedy this problem, we still cannot guarantee which method is better than others. Therefore, the very next step is to collect more individual calling samples and validate the system thoroughly.

The MLP classifier used can only classify those already known classes and demands a long training time before testing. In the near future, the proposed frog call monitoring system is required to be placed in field for real time monitoring. It will not have any prior knowledge about any individual frog call pattern and neither will it know how many individuals are calling during the recording period. In this case the classifier must be built in with on-line learning ability that can learn new patterns in real-time and continuously grow the number of identified individual numbers without re-training. MLP is clearly defeated by this specification. The Incremental Learning Fuzzy Neural Network (ILFN) [20] can address this deficiency. It uses an incremental learning algorithm and can detect new classes of patterns and update its parameters while in an operating mode. And it has an on-line (real-time) and fast learning algorithm without knowing a-priori

information. Later, after enough data had been acquired, this classifier may be explored further. It may detect new individual frogs in real time and there is no need to train the network with known features.

High-level environmental noise may unnecessarily complicate the identification process than the simulations conducted in laboratory environments. How to realize noise cancellation or noise reduction, especially for those noises with similar frequency range as frog calls, remains a challenging issue.

## REFERENCES

1. Bantle, J., Private Communication. 1999.
2. Breiman, L., "Comment", added to "Neural Networks: A Review from a Statistical Perspective," by Cheng, B., and Titterington, D. M., Statistical Society, Vol.9, pp. 2-54, 1994.
3. Deller, J. R., Proakis, J. G. and Hansen, J. H. L., Discrete-Time Processing of Speech Signals, Macmillan, New York, NY, 1993.
4. Friendly, M, "Bartlett's Test for Homogeneity of Variances: A SAS Macro," Department of Psychology, York University, Toronto, Ontario, Canada, 1995; <http://www.math.yorku.ca/SAS/friendly.html>.
5. Fristrup, K. M. and Watkins, W. A., "Marine Animal Sound Classification," Journal of the Acoustical Society of America, Vol. 97, No. 5, pp. 3369-3370, May 1995.
6. Gerhardt, H. C., "Acoustic Properties Used in Call Recognition by Frogs and Toads," In Fritzsche *et al.*, editor, The Evolution of the Amphibian Auditory System, pp. 455-483, John Wiley, New York, NY, 1988.
7. Haykin, S., Neural Networks: A Comprehensive Foundation, Macmillan, New York, NY, 1994.
8. John, G. H., Kohavi, R. and Pflieger, K., "Irrelevant Features and the Subset Selection Problem," Proceedings of the 11<sup>th</sup> International Conference on Machine Learning, pp. 121-129, 1994.
9. Kientzle, T., A Programmer's Guide to Sound, Addison-Wesley, Reading, MA, 1998.
10. Kittler, J., "Mathematical Methods of Feature Selection in Pattern Recognition," International Journal on Man-Machine Studies, Vol. 7, pp. 609-637, 1975.
11. Lieberman, P. and Blumstein, S. E., Speech Physiology, speech Perception and Acoustic Phonetics, Cambridge University Press, Cambridge, UK, 1988.
12. Lin, Z. B., "Some Aspects of Wavelet Transform in Bio Sonar Processing and Detection," Journal of the Acoustical Society of America, Vol. 91, pp. 2468-2468, 1992.

13. Lippmann, R. P., "Review of Neural Networks for Speech Recognition," Neural Computation, Vol. 1, pp. 1-38, 1989.
14. Mallat, S. G., "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 11, No. 7, pp. 674-693, July 1989.
15. Markel, J. D. and Gray, A. H., Linear Prediction of Speech, Springer-Verlag, Berlin, New York, 1976.
16. Matson, T. O., "An Introduction to the Natural History of the Frogs and Toads of Ohio," <http://www.cmnh.org/research/vertzoo/frogs/>.
17. McIlraith, A. L. and Card, H. C., "Birdsong Recognition with DSP and Neural Networks," Proceedings of IEEE Conference on Communications, Power, and Computing, Vol 2, pp. 409-414, 1995.
18. McIlraith, A. L. and Card, H. C., "Birdsong Recognition Using Backpropagation and Multivariate Statistics," IEEE Transactions on Signal Processing, Vol. 45, No. 11, pp. 2740-2748, November 1997.
19. McLelland J. L. and Rumelhart, D. E., Explorations in Parallel Distributed Processing, MIT Press, Cambridge, MA, 1989.
20. Meesad, P., Pattern Classification by An Incremental Learning Fuzzy Neural Network, Master thesis, School of Electrical and Computer Engineering, Oklahoma State University, 1998.
21. Mendel, J. M., Lessons in Estimation Theory for Signal Processing, Communications, and Control, Prentice Hall, Englewood Cliffs, NJ, 1995.
22. Mellinger, D. K. and Clark, C. W., "Methods for Automatic Detection of Mysticete Sounds," Marine and Freshwater Behavior and Physiology, Vol. 29, pp. 163-181, 1997.
23. Mills, H., "Automatic Detection and Classification of Nocturnal Migrant Bird Calls," Journal of the Acoustical Society of America, Vol. 97, No. 5, pp. 3370, May 1995.
24. Misiti, M., Misiti Y., Oppenheim, G., and Poggi, J. M., Wavelet Toolbox User's Guide, the Math Works, Inc., Natick, MA, 1996.
25. Neter, J. and Wasserman, W., Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Designs, R. D. Irwin, Homewood, IL, 1974.
26. Oppenheim, A. V. and Schaffer, R. W., Discrete-Time Signal Processing, Prentice Hall, Upper Saddle River, NJ, 1998.

27. Osborne, W. S., "Frogs," Microsoft® Encarta® Online Encyclopedia 2000, 2000, <http://encarta.msn.com>.
28. Pickles, J. O., An Introduction to the Physiology of Hearing, Academic, London, 1988.
29. Pielou, E. C., The Interpretation of Ecological Data: A Primer on Classification and Ordination, Wiley, New York, NY, 1984.
30. Potter, J. R., Mellinger, D. K. and Clark, C. W., "Marine Mammal Call Discrimination Using Artificial Neural Networks," Journal of the Acoustical Society of America, Vol. 96, No. 3, pp. 1255-1262, September 1994.
31. Quinlan, J. R., C4.5: Programs for Machine Learning, Morgan Kauffman, 1993.
32. Rabiner, L. R. and Gold, B., Theory and Application of Digital Signal Processing, Prentice Hall, Englewood Cliffs, NJ, 1975.
33. Rabiner, L. R. and Juang, B. H., Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, NJ, 1993.
34. Riedmiller, M., and Braun, H., "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm," Proceedings of the IEEE International Conference on Neural Networks, Vol.1, pp. 586-591, 1993.
35. Sasaki, K. and Yamazaki, M., "Vector Compression of Bird Songs Spectra in Water Sites by Using the Linear Prediction Method and its Application to An Automated Bayesian Species Classification," 38th Annual Conference Proceedings of the SICE Annual, pp. 1083-1088, 1999.
36. Squires, B. and Sammut, C., "Automatic Speaker Recognition: An Application of Machine Learning," Proceeding of the 12<sup>th</sup> International Conference on Machine Learning, 1995.
37. Strang, G., and Nguyen, T., Wavelets and Filter Banks, Wellesley-Cambridge Press, Wellesley, MA, 1996.
38. Taylor, A. J., "Bird Flight Call Discrimination Using Machine Learning," Journal of the Acoustical Society of America, Vol. 97, No. 5, pp. 3370, May 1995.
39. Taylor, A. J., Watson, G., Grigg, G., C., and McCallum, H. I., "Monitoring Frog Communities: An Application of Machine Learning," Proceedings of the 8th Innovative Applications of Artificial Intelligence Conference, 1996.

40. Yen, G. G. and Lin, K. C., "Wavelet Packet Feature Extraction for Vibration Monitoring," IEEE Transactions on Industrial Electronics, Vol. 47, No. 3, pp. 650-667, June 2000.
41. Fukunaga, K., Introduction to Statistical Pattern Recognition, Academic Press, Inc., 1992.

VITA

Qiang Fu

Candidate for the Degree of

Master of Science

Thesis: Automatic Frog Calls Monitoring: A Machine Learning Approach

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Wuxi, China, On November 7, 1975, the son of Guoping Fu and Meifeng Sheng.

Education: Received a Bachelor of Science in automation degree from Tsinghua University, Beijing, China in June 1998. Completed the requirements for the Master of Science degree with a major in Electrical Engineering at Oklahoma State University in December, 2000.

Experience: Employed as an electrical engineer, 1998 by HengDa Co., Ltd in Hefei; employed by Oklahoma State University, School of Electrical and Computer Engineering as a research assistant, 1999 to 2000.