AUTOMATIC CLASSIFICATION OF ANIMAL VOCALIZATIONS

by

Patrick J. Clemins, B.S., M.S.

A Dissertation submitted to the Faculty of the
Graduate School, Marquette University,
in Partial Fulfillment of
the Requirements for
the Degree of
Doctor of Philosophy

Milwaukee, Wisconsin

May, 2005

PREFACE

At the beginning of the doctoral process, I knew I had two important decisions to make. The first was who would be my research advisor and the second was what would be the topic of my dissertation. I had not spent much time on either decision during my master's degree, and although I am happy with the way things transpired, I had set higher goals for my doctoral research. I found a wonderful complement to myself in Michael T. Johnson, Ph.D., a new faculty member from Purdue's doctoral program and knew right off that he would make my doctoral process challenging and rewarding. However, I did not know at first that the process would also be enjoyable and exciting.

After I had picked an advisor, I needed to pick a topic on which to dissertate. A new topic on the leading edge of a research field that I could easily explain to people, mainly my family and friends, would be the ideal. I did not want to work in an older, well-established field where most current contributions were incremental improvements. These incremental improvements are important, however, I wanted to do something different and unique. Then, Mike and his wife Patricia took a vacation to Disney for their wedding anniversary. While he was there, he noticed an exhibit that explained how the researchers were recording elephants to study their vocal communication. After some coaxing by Patricia, Mike asked for some contact information for the researchers collecting and analyzing the elephant vocalizations. It was then that he met Dr. Anne Savage, and a collaboration was formed. He came back to Marquette and asked me if I wanted to work on bioacoustics. After an explanation of what exactly bioacoustics was, I knew I had found my topic. It was on the leading edge of human knowledge about our world, and yet most people could relate to the desire to know what animals were trying to communicate vocally.

My role in this project has been to show that the application of speech processing techniques to animal vocalization is viable and to establish a framework that can be adapted to multiple species. The results achieved thus far are much better than anticipated. Once the feature extraction model was modified to incorporate information about each species' acoustic perception abilities, the results were even better, and a robust and flexible framework was established.

The project as a whole has been stimulating, not only because the field is new, but also because of its multidisciplinary nature. The understanding of the vocal communication structure of non-human species requires the expertise of biologists to understand the physical abilities of each species, psychologists and neurologists to understand the processing of sound in the brain, animal behaviorists to determine what the animal was trying to communicate, engineers to design recording equipment and process data, and many others. I am grateful that many people are interested in this goal of trying to better understand animals to learn how to better share this planet with them and improve conservation efforts.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support of numerous colleagues, family members, and friends. Thank you to my committee for all of the insightful comments and encouragement along the way. Thanks to all my colleagues and friends for suggestions and especially Ricardo Santiago for proofreading the entire document. Thank you to the staff of the Wildlife Tracking Center and Elephant Team at Disney's Animal Kingdom™, especially Anne Savage, Kirsten Leong, and Joseph Soltis for all the hard work collecting, organizing, and sharing their African elephant data. Speaking of data, thank you also to Pete Scheifele for the beluga data, the wonderful conversations, and the colorful insights and humor that provided a nice break when needed. Thank you also to the ASA bioacoustic community for their open-mindedness and well-structured research community. I am appreciative of all of the contacts I have made at the various meetings.

Thanks to my parents and family for their support and understanding in my pursuit of knowledge. Thank you also to all my friends in Triangle, volleyball, TFB, bible study, from Mad Planet, and others, who were always there to listen and provide some much needed downtime. Finally, thanks to God for providing us with this wonderful and diverse world in which to live, experience, and discover.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Purpose

Bioacoustics, the study of animal vocalizations, has recently started to explore ways to automatically detect and classify vocalizations from recordings. Although there have been a number of successful systems built for specific vocalizations of a particular species, each is built using different models for quantifying the vocalization as well as different classification models. The purpose of the research is to develop a generalized framework for the analysis and classification of animal vocalizations that can be applied across a large number of species and vocalization types. Such a framework would also allow researchers to compare their classification results with those of other researchers fairly.

The framework applies popular techniques found in current state-of-the-art human speech processing systems and is based on the perceptual linear prediction (PLP) feature extraction model (Hermansky, 1990) and hidden Markov model (HMM) classification model. Because the framework incorporates a generalized version of the PLP feature extraction model, the framework is called the generalized perceptual linear prediction (gPLP) framework. These techniques are modified to compensate for the differences in the perceptual abilities of each species and the structure of each species' vocalizations. The modifications are made to incorporate various amounts of available knowledge of the species' perceptual abilities and vocalization structure. Therefore, the modifications can be applied to varying degrees depending on perceptual information available about the species being analyzed.

One drawback about traditional bioacoustics signal analysis is that it calculates statistics once over the entire vocalization. For highly dynamic vocalizations, statistics calculated over the entire vocalization do not adequately describe the time-varying nature of the vocalization. Even though statistics can be designed to measure the dynamics of a vocalization such as the slope of the fundamental frequency contour or the minimum and maximum fundamental frequency, the number of statistics needed to accurately model complex vocalizations soon becomes a hindrance. It is also difficult to compare vocalizations of different types using special dynamic statistics since each type of vocalization may require a number of these dynamic statistics, and some statistics may not be applicable to certain vocalizations.

The gPLP framework captures vocalization dynamics by breaking the vocalization into windows, or frames, and then calculates a number of values which quantify the vocalization every frame. These calculated values are called features and the set of values calculated each frame is known as the feature vector. It is assumed that the signal is approximately stationary for the duration of the frame. For a signal to be stationary, the spectral characteristics cannot change for the duration of the signal. Accurate spectral estimation depends on the signal being stationary for the entire analysis window. Since the spectral characteristics of a vocalization are constantly changing, the vocalization is framed to calculate accurate spectral estimates.

The process of dividing the vocalization into frames and generating a feature vector is analogous to the process for generating a spectrogram. In a spectrogram, the spectrum, the feature vector in this case, is calculated over windows and then plotted versus time. Therefore, the vocalization is quantified into a matrix of features calculated each frame instead of a vector of single feature values calculated over the entire vocalization which is

common in traditional bioacoustics. A typical feature matrix is shown below in Equation 1.1.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,C} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,C} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,C} \end{bmatrix}, \quad 1.1$$

where C is the number of features and N is the number of frames in the vocalization.

The other major change from historical bioacoustics analysis is that the vocalizations are modeled with a Hidden Markov Model (HMM). Although described in more detail in Chapter 2, an HMM is a statistical model that can model time variations in the vocalizations. The extraction of features on a frame-by-frame basis requires a more complex model than a single vector of features measured over the entire vocalization to deal with the additional data. The HMM breaks the vocalization into a number of distinct states and assigns each frame of data to a state. The HMM is a good choice because it can model both the temporal and spectral differences in a vocalization and perform non-linear time alignment of different examples of a vocalization during both training and classification.

Motivation

Bioacoustics traditionally relies on spectrogram analysis to generate features. Typical features are fundamental frequency measures and duration values. Many of these values are measured through visual inspection and are thus susceptible to researcher bias. These similar measures also tend to be highly correlated with each other and therefore are not the most appropriate set of features for a statistical classifier. The gPLP framework outlined in this dissertation includes a feature calculation, or extraction, model which uses information about how an animal perceives sound as its basis. An automatic feature extraction model can generate unbiased features while drastically reducing the time spent measuring them.

Statistical tests, commonly used in bioacoustics research, although adequate for validating scientific hypotheses, are not designed for classification tasks. A complete classification model has well-developed methods for training the model as well as a method for classifying new data. The HMM is particularly suited for classifying time series such as vocalization waveforms due to its ability to model non-linear temporal variations in the vocalizations. An automatic classification system, consisting of a feature extraction model and a classification model, can provide a robust and efficient way to analyze animal vocalizations.

Automatic classification systems can help uncover acoustic patterns related to the psychological or physiological state of the animal that are not obvious from examining spectrograms which do not incorporate perceptual information. The incorporation of traditional bioacoustics methods with an automatic classification system makes it possible to not only test hypotheses, but also build systems which use these hypotheses to classify unknown vocalizations, find new vocalizations, and measure how much vocalizations vary between and within each type or class.

This research is part of the Dr. Dolittle Project, a multi-institution research effort funded by the National Science Foundation. The goal of the project is to apply automatic classification systems to bioacoustic tasks. Automatic classification systems take a raw signal and assign a class label without human intervention. Automatic classification systems usually consist of a feature extraction model which describes how to quantify the signal and a classification model which describes how to compare vocalizations to each other and assign a class label to the signal. In addition to the research presented in this dissertation, there are ongoing projects in signal enhancement, incorporating seismic data, and building classification systems for avian species. The work presented in this dissertation represents

the majority of the preliminary work for the project as well as a generalized feature extraction model that can be applied to the various species being studied.

Current Trends

This research fits in well with current trends in bioacoustics and speech recognition research. Speech recognition, the machine translation of speech to text, has been an active area of speech research for the last thirty years. One current branch of speech recognition research involves creating viable speech recognition systems (Picone, 1990; Roe and Wilpon, 1993) which can be put into practical use. These systems have been successful due to the use of improved acoustic features, more advanced language models, and adaptation to the domain in which the system will operate. For instance, systems that are expected to operate in high noise conditions are either trained with data corrupted with appropriate domain-specific noise or equipped with a noise-reduction front-end processor. The gPLP framework developed in this dissertation addresses two of these techniques, improved acoustic features, and adaptation to the domain. Improved acoustic features are realized by integrating information about the animal's perceptual abilities, and the system is adapted to the domain through the topological design of the HMMs.

Current trends in bioacoustics signal analysis are to incorporate automatic methods for classification and feature extraction (Buck and Tyack, 1993; Campbell *et al.*, 2002; Mellinger, 2002; Weisburn *et al.*, 1993). However, no common framework for these processes has been established. Therefore, each experiment uses different features and different types of classification systems, making inter-species comparison studies extremely difficult. In addition, it is hard for researchers to build upon other systems since each is designed for a particular species and domain. This research hopes to begin to establish a framework that

can help to unify the research community and show that the framework is viable across different species and domains.

Applicability

Speaker identification, speech recognition and word-spotting, common tasks in speech processing, directly relate to tasks in bioacoustics. Speaker identification, the determination of the individual speaking, can be used to help label bioacoustic data by determining which animal is speaking even though the animal is out of sight. It is also a vital component in the creation of a census application. An automatic census system, consisting of a set of microphones deployed in the animal's natural habitat and a software package, could estimate the local population of the species using the number of unique speakers in the area of deployment. Speech recognition, the translation of speech to text, is analogous to determining the meaning of a vocalization or determining the type of a vocalization. Word-spotting, the detection of specific words in a conversation, can detect animal vocalizations in a lengthy recording and divide the vocalizations into individual signals. The process of extracting individual vocalizations from a recording session is known as segmentation in the speech processing field.

Speech processing techniques are attractive because of the large amount of effort given to the field over the last fifty years. Speech systems incorporate feature extraction techniques that are robust to noise with optimal statistical classification models. Recent research has shown that other animals' vocal production and perception mechanisms can be modeled by the source-filter model which is the basis behind human speech systems (Bradbury and Vehrencamp, 1998; Fitch, 2003; Titze, 1994). Therefore, it is reasonable to hypothesize that human speech algorithms can be adapted to other species since the production and perception mechanisms are similar.

Main Challenges

The major challenges in adapting human speech processing algorithms to bioacoustics include background noise, lack of specific knowledge about animal communication, and label validity. First, the background noise conditions are not easily controlled when collecting animal vocalizations. Speech systems trained and evaluated on data collected in controlled environments are much more successful than those evaluated in real-world environments where changes in recording conditions are inevitable. There are techniques to reduce the effects of these mismatched conditions, but they usually do not eliminate the effects of noise. It is difficult to train an animal to vocalize naturally in an acoustically controlled room, thus most vocalizations are recorded in naturalistic conditions or in the wild where traffic noise and other animal vocalizations can interfere with the data collection.

Although we have a good understanding of the important perceptual features in human speech, it is not clear what components of the vocalization hold meaning for intra-species communication. Although the pitch of the vocalization is only important linguistically in a few tonal human languages, it seems to be much more important in many species' communication based on the number of studies that include pitch as a feature, sometimes as the only feature (Buck and Tyack, 1993; Darden *et al.*, 2003). In addition, human speech processing systems primarily use spectral features, but there is no reason to rule out temporal features such as amplitude modulation or number of repetitions of a syllable as the important components of an animal vocalization. Finally, once these salient features are determined, automatic extraction algorithms need to be created to extract them to make a fully automatic classification system.

Another problem is the validity of the labels given to bioacoustic data. In speech experiments, sentences can be given to subjects. In this case, the subject and researcher

both know the intended meaning of the vocalization. Even in unconstrained speech experiments, researchers have full knowledge of our language and can interpret the meaning correctly. This knowledge is not present or minimal for animal vocalizations. Behavior can be one cue toward the purpose of a vocalization, but the cues are often ambiguous. Even labels such as the individual making the vocalization can be difficult to determine if the recording is made in an environment such as a forest where line-of-sight of the speaker may not be available.

One last difficulty is that the physiology of how animals make and perceive vocalizations is not always known. Although numerous experiments have been performed on the human auditory and speech system, much information is still lacking for many species. Without basic experimental data or information about the mechanisms behind sound generation, it is extremely difficult to model both the production and perception of sound for that species.

Contributions

The development of a generalized analysis framework for animal vocalizations that can be applied to a variety of species is the main contribution of this research. The generalized perceptual linear predication (gPLP) extraction model, which incorporates perceptual information for each species, contains the majority of the novel ideas. The application of gPLP features to statistical hypothesis testing, supervised classification, and unsupervised classification demonstrates some practical uses of the gPLP framework to bioacoustic tasks. The signal processing involved in creating the gPLP extraction model contributes to the field of electrical engineering, while the application of human speech features and models to animal vocalizations is the most important contribution to the field of bioacoustics.

Dissertation Overview

The first chapter has been a brief overview of the dissertation and the motivation behind the research. The second and third chapters discuss the necessary background knowledge in the fields of speech processing and bioacoustics respectively.

The fourth chapter details the gPLP framework and highlights the modifications that were made to traditional speech processing techniques when applied to animal vocalizations. It also highlights the preliminary research that led to the development of the gPLP framework.

The fifth and sixth chapter applies the gPLP framework to two different classification tasks. The first is a supervised classification task on elephant vocalizations, and the second is an unsupervised classification task using beluga whale vocalizations. The final chapter, seven, gives a summary of the dissertation, discusses the contributions of the research, and suggests possibilities for future work.

Chapter 2

SPEECH PROCESSING

Background

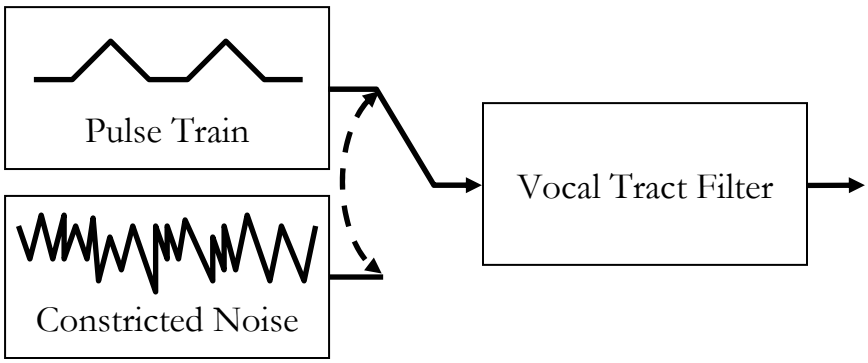
The field of speech processing, with its roots in attempts to understand speech production (Deller *et al.*, 1993), has been firmly established for over fifty years. It embraces research in many fields including linguistics, signal processing, speech pathology, psychology, neurology, physiology, and others. Current speech processing systems incorporate feature extraction algorithms robust to background noise and optimal statistical classification techniques. The two areas of speech research with which this dissertation is most closely associated are speech recognition, the conversion of spoken speech into written text, and speaker identification, the determination of the individual speaking.

The traditional speech analysis process is based on the source-filter model of speech production shown in Figure 2.1 (Dudley, 1940; Flanagan, 1965). This relationship can be represented mathematically by

$$y[n] = s[n] \otimes h[n], \tag{2.1}$$

where $\otimes$ is the convoluation operator, $y[n]$ is the vocalization, $s[n]$ is the excitation, and $h[n]$

Figure 2.1 – Source-Filter Model of Speech



is the vocal tract filter which is assumed to be linear.

In this model, the glottis, also called the vocal folds, generates an excitation signal. The excitation is then filtered by our vocal tract and articulators. The articulators are the parts of the vocal tract which are actively controlled to generate difference types of sounds, namely the tongue, teeth and lips. Human speech is generally separated into two types of sounds, voiced and unvoiced, based on the excitation signal. Voiced speech is the result of a pulse train excitation signal, while unvoiced speech is the result of a white noise excitation signal. One example of this difference is the unvoiced 's' in sip and the voiced 'z' in zip. The difference in excitation can be felt by placing fingers on the throat while vocalizing the two different words. The greater vibration of the voiced 'z' can be clearly felt.

Voiced speech is produced as a result of closing the vocal folds by tightening the muscles around them. The air pressure then builds up below the folds until the pressure forces them open. A Bernoulli force then pulls the folds closed, and the pressure builds up for the next air pulse. This sequence of air pulses generates a pulse train, the excitation signal for voiced speech. Unvoiced speech is produced by allowing the vocal folds to be open and relaxed. This causes the vocal folds to vibrate as air is forced through the gap, generating a white noise signal.

Active manipulation of the articulators allows for the production of different types of sounds. Each different sound is called a phoneme, the basic unit of sound in human speech. The English language has about 50 phonemes which have been identified in various phonetic alphabets. One of the more common phonetic alphabets in use today is ARPAbet, developed under the Advanced Research Projects Agency (Deller *et al.*, 1993). Its popularity stems from the fact that the symbols ARPAnet uses for phonemes consist of only ASCII characters. For example, the word fish, although consisting of four letters, has three

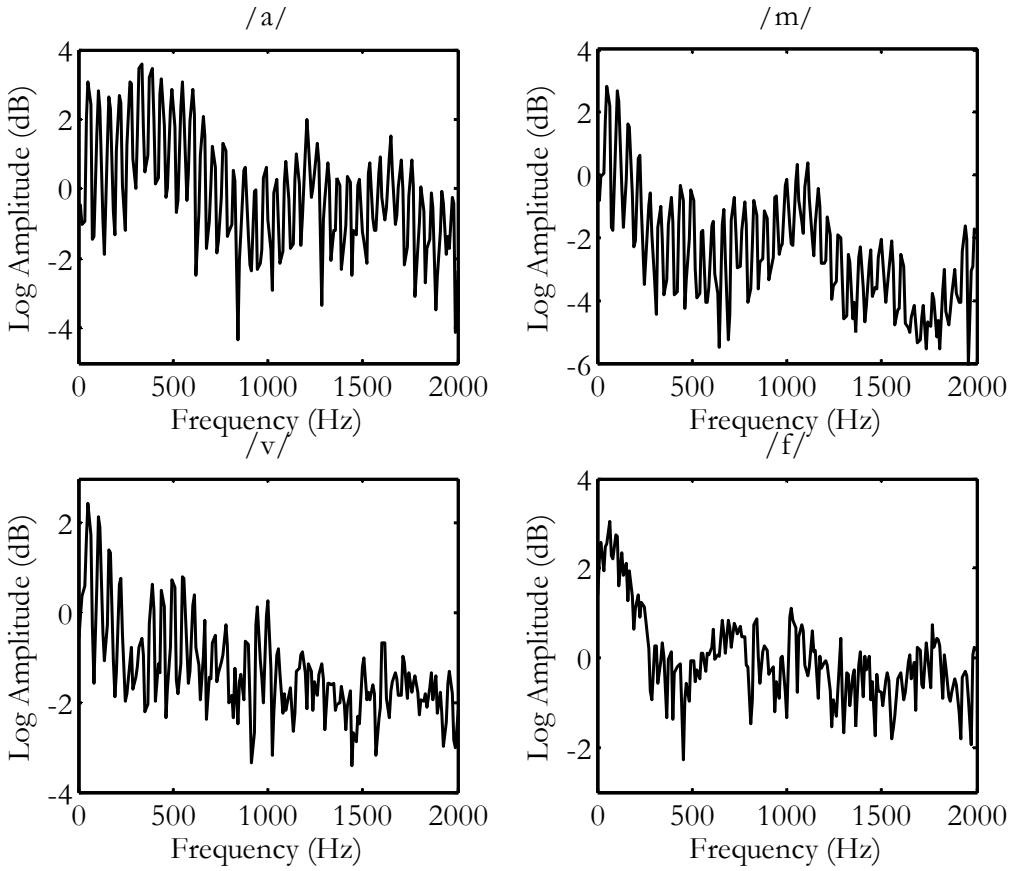


Figure 2.2 – Frequency Spectra for Various Phonemes

phonemes, /f/, /I/, and /S/. The voicing (i.e. voiced or unvoiced) of a phoneme and the position of the articulators together uniquely define a phoneme in English and most other human languages. For instance, the phonemes /p/ as in 'pig' and /b/ as in 'big' have the same articulator structure, but /p/ is unvoiced, while /b/ is voiced. Other languages such as Thai and Mandarin add tone, or pitch, as a unique identifier to the phoneme. In these tonal languages, the pitch contour, along with voicing and articulator position, uniquely define a phoneme.

The position of the articulators is manifested in the frequency spectrum by peaks in the spectral envelope, while the voicing is evident in the whether the envelope is jagged or sinusoidal. The peaks in the spectral envelope are called formants. Three to four formants are typically visible in human speech spectra, and their positions quantify the shape of the

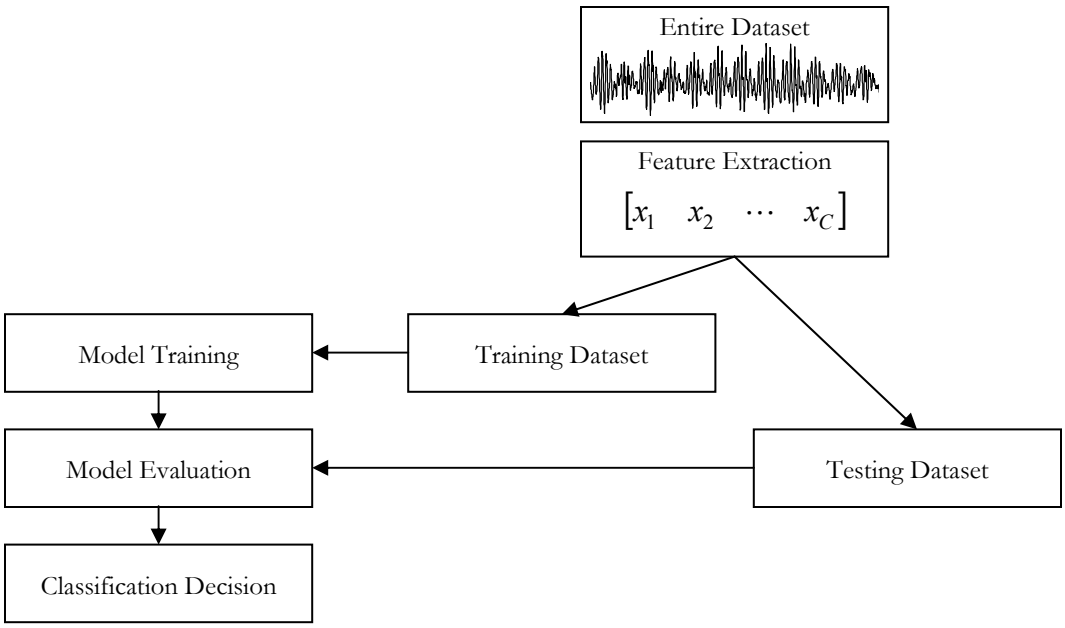


Figure 2.3 – Speech Processing Classification System

vocal tract filter as described in the source-filter model. Four example spectra are in Figure 2.2. The bottom two spectra, from the phonemes /v/ as in 'voice' and /f/ as in 'fish', both have formant locations at about 500Hz and 1000Hz. However, one is voiced, while the other is unvoiced. This difference in voicing is manifested in the spectra by the sinusoidal envelope of the voiced phoneme, /v/, and the jagged envelope of the unvoiced phoneme, /f/. The other two examples show a vowel, /a/ as in 'soda', and a nasal, /m/ as in 'mouse' to compare the different formant locations of each phoneme. The phoneme /a/ has formants at 500Hz, 1250Hz and 1750Hz, while the phoneme /m/ has formants at 1100Hz, 1500Hz and 2000Hz.

To accomplish the tasks of speech recognition and speaker identification, current speech processing methods include a feature extraction front end followed by a classification system which is usually statistical. A block diagram of this process is shown in Figure 2.3. The original speech waveform is not analyzed, but instead, features extracted from the vocalization are modeled by the classification system. Current feature extraction algorithms

are based on frequency analysis. Features are calculated in the spectral domain using a Fourier transform to estimate the spectrum (Hunt, 1999). The speech is broken up into short 10ms – 30ms frames to keep the signal stationary for the spectral analysis. For a signal to be stationary, the spectral characteristics of the signal must not change. If the signal is not stationary for the duration of the analysis window, the spectral estimate will lose accuracy. The features calculated in each frame are then concatenated together to generate a feature matrix as in equation 1.1. This feature matrix is then used as input into a classifier.

Although the salient components of speech are best described in the spectral domain, the cepstral domain has become the preferred domain of speech features because of its beneficial mathematical properties (Deller *et al.*, 1993). The name is derived from swapping the consonant order in the word spectral, the domain from which the cepstral domain is derived. Similarly, filters in the cepstral domain are commonly called lifters, and the plot of cepstral values is called the cepstrum, the complement to the spectrum. The cepstral domain is the inverse Fourier transform of the logarithm of the Fourier transform of a signal. Mathematically this is represented by

$$C[m] = F^{-1}(\log(F(s[n]))), \tag{2.2}$$

where F is the Fourier transform, and $s[n]$ is the original discrete time domain signal. The cepstral domain has become popular because the general shape of the signal and spectrum can be accurately represented by a small number of cepstral coefficients. This representation is advantageous from a source-filter model perspective as well. While the excitation source is convolved with the vocal tract filter in the time domain, it is a simple addition in the cepstral domain as represented by

$$\log Y[m] = \log S[m] + \log H[m]. \tag{2.3}$$

Therefore, the source, $S[m]$, can be easily separated from the filter, $H[m]$, through subtraction. The logarithm operation also separates the filter and excitation in the cepstrum. The first few cepstral coefficients represent the slow moving components of the spectrum, the filter, while the excitation appears as a triangular peak in higher indexed cepstral coefficients. Other filters or convolutional noise which may be part of the speech signal, such as channel distortion, can also be subtracted from the cepstrum.

Common Tasks

The two tasks in speech research that are most associated with this research are speech recognition and speaker identification. Although there are many more applications of speech research, they will not be discussed here.

Speech Recognition

Speech recognition is the translation of acoustic data to written text. Many of the first systems were isolated word recognition systems. These systems required the words to be pre-segmented into separate signals and there was one classification model trained for each word in the vocabulary. However, segmentation errors and the number of models required for a large vocabulary made these types of systems difficult to design. Therefore, most current systems are statistical-based, continuous recognition systems using models based on phonemes and incorporate dictionaries of word pronunciations and language models to guide the recognition process (Rabiner and Juang, 1993).

Speech recognition systems are becoming more widely implemented, especially in customer care call centers, where the user can speak into the phone instead of choosing numbers for menu items. It is also being used more in the personal computing environment as a user interface and has been included in some of the more recent operating systems and applications, especially those made by Microsoft®. Single speaker systems with medium

vocabularies in a low-noise environment can approach word accuracies of 92% (Padmanabhan and Picheny, 2002). Current research topics in this area include robust speech recognition in the presence of noise and model adaptation for different recording environments.

Speaker Identification

Speaker identification is the determination of the speaker of a segment of acoustic data (Campbell Jr., 1997). Its most common application is in biometrics and security systems. To determine the speaker, the unknown utterance is compared to a number of speaker models trained on speech from each speaker. The speaker's model that most closely matches the unknown speech is the hypothesized speaker. The task can be either closed set or open set. In a closed set speaker identification task, the system is forced to pick a speaker from the database of known speakers. However, in an open set task, the system may decide that the speaker of the test utterance is not in the database of known speakers and is instead an unknown speaker. The task can also be defined over the same phrase or over unspecified speech. When the phrase is the same, the task is called text-dependent; when the phrase is unspecified, it is referred to as a text-independent task.

Speaker identification systems have been implemented commercially for user verification purposes. Current state-of-the-art systems can reach recognition rates of 85% on telephone-quality speech with thousands of possible speakers (Reynolds, 2002). The accuracy decreases with lower-quality speech and a larger speaker database. Current research topics in this area include channel normalization, training new speaker models with small amounts of data, and quick lookup techniques for large speaker databases.

Feature Extraction

The features extracted from the vocalization waveform are very important to the success of a classification system. Features that are sensitive to noise, susceptible to bias, or do not discriminate between classes only confuse the system and decrease classification accuracy. Ideally, features should be unbiased, uncorrelated, and reflect the characteristic differences between classes. In speech, spectral-based features have been the most commonly used.

The two most popular features currently used in speech processing research are Mel-Frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Prediction (PLP) coefficients. MFCCs were originally developed by Davis and Mermelstein (1980). They are the most popular features because of their computational efficiency and resilience to noise. The ability of MFCCs to capture vocal tract resonances but exclude excitation patterns and tendency to be uncorrelated are also beneficial characteristics.

The PLP model was developed by Hermansky (1990) more recently and stresses perceptual accuracy over computational efficiency. Hermansky showed that PLP coefficients are also resistant to noise corruption and demostrate some of the same responses to noise that the human auditory system exhibits. Although MFCCs are still more commonly used, PLP analysis is gaining in popularity and is more suited for this research because it allows more information about the sound perception system to be incorporated into the model (Milner, 2002).

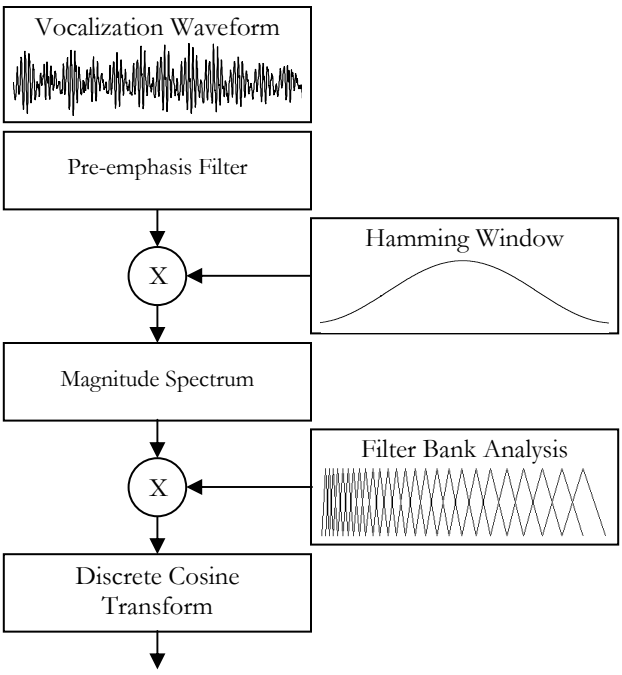


Figure 2.4 – MFCC Block Diagram

MFCC Feature Extraction

The MFCC feature extraction model as described by Davis and Mermelstein (1980) is outlined in Figure 2.4. Each part of the block diagram is explained in the sections below.

Pre-Emphasis Filter

The feature extraction process begins by applying a time-domain pre-emphasis filter to the vocalization, $s[n]$. The purpose of the pre-emphasis filter is two-fold (Deller *et al.*, 1993). First, the pre-emphasis filter tends to cancel out the effects of the larynx and the lips on the vocal tract filter. This is desirable because the position of the larynx and lips do not contribute much information about the phoneme being uttered. Second, the pre-emphasis filter helps to compensate for spectral tilt. Spectral tilt is the tendency for the spectral envelope to gradually decrease in value as frequency increases. In the case of human speech, it is caused by general nature of the human vocal tract. Spectral tilt increases the dynamic range of the spectrum. This increased dynamic range forces the discrete cepstral transform,

which occurs later in the feature extraction process, to focus on the larger peaks in the spectrum occuring in the lower frequencies. The pre-emphasis filter decreases the dynamic range by emphasizing the upper frequencies and suppressing the lower frequencies. As a result, the discrete cosine transform can model the higher frequency formants with better consistency. The filter is of the form

$$s'[n] = s[n] - \alpha s[n-1], \tag{2.4}$$

with α having a typical value of about 0.97.

Hamming Window

The next step is to break the vocalization into frames and multiply each frame by an analysis window to reduce artifacts that result from applying a spectral transform to finite sized frames. Human speech processing typically uses a Hamming window defined by

$$w[n] = 0.54 + 0.46\cos\left[2\pi n \big/ (N-1)\right], \tag{2.5}$$

where N is the length of the frame. Another windowing function commonly used, especially in bioacoustics analysis, is the Hanning window. Both windows are extremely similar in structure (Oppenheim and Schafer, 1999:468).

Magnitude Spectrum

A spectral estimation method, typically a Fourier transform, is then applied to the windowed frame to acquire a magnitude spectrum. Although many different techniques could be used to generate a spectral estimate, a Fast Fourier Transform (FFT) is the most common approach.

The length of the analysis window, and consequently the FFT, has a large effect on the time and frequency resolution of the spectrum. A larger frame size results in improved frequency resolution, but because it requires more time samples, the time resolution

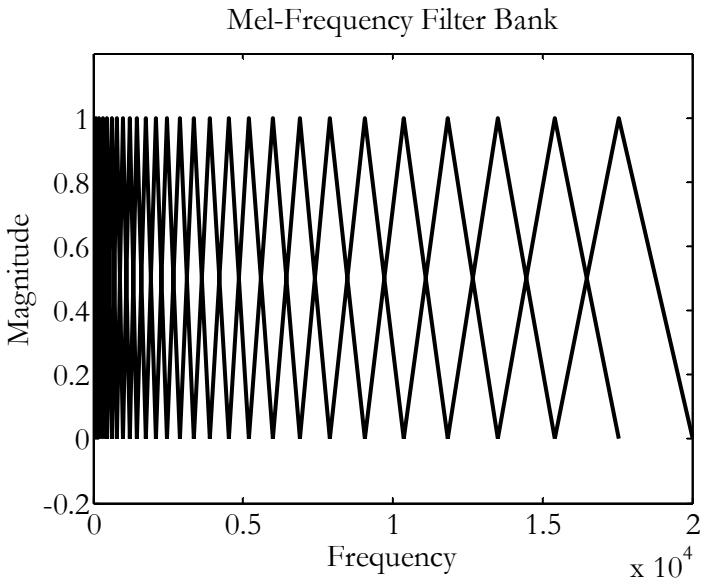


Figure 2.5 – Mel-Frequency Filter Bank

decreases. This can be partially compensated by increasing the frame overlap, however too much overlap results in data being reused and duplicated in multiple frames. Larger frame sizes also tend to decrease the stationarity of the signal and thus reduce the accuracy of the spectral estimate. A shorter frame size improves the temporal resolution at the cost of frequency resolution since the frequency resolution is inversely proportional to the window size.

Filter Bank Analysis

The next step in the MFCC model is to perform filter bank analysis on the magnitude spectrum. A filter bank is simply a set of filters as shown in Figure 2.5. Here, the main purpose of the filter bank is to model human psychoacoustics and accurately represent human perception of speech. The filter bank takes into account logarithmic sensitivity to frequency and frequency masking. Stevens and Volkmann (1940) showed that humans perceive sound on a logarithmic scale in reference to cochlear position sensitivity. The most popular approximation of this scale is the Mel-scale because it has a mathematically simple representation. The Mel-scale is defined as

$$f_{Mel} = 2595 \log_{10}(1 + f/700). \quad 2.6$$

Frequency masking occurs when a signal has spectral peaks near each other. The stronger frequency effectively cancels out weaker frequencies around it. Consequently, only the stronger frequency is perceived. The smallest change in frequency that can be perceived is called the critical band. The critical band is a function of frequency, increasing as frequency increases. This masking effect can be modeled by a critical band masking curve which suppresses surrounding frequencies, making them harder to perceive. A triangular filter shape approximates the critical band masking curve as determined experimentally (Glasberg and Moore, 1990; Patterson, 1976).

MFCC analysis uses Mel-frequency spaced triangular shaped filters. A diagram of Mel-frequency spaced filters is shown in Figure 2.5. The number of filters, usually about 26, is determined by the number of critical bands (Zwicker *et al.*, 1957) that can be laid out side-by-side across the human hearing range. The spectral energy in each filter band is totaled by multiplying each filter by the spectrum and summing the result. A sequence of filter bank energies is generated which correlates with a low-pass filtered, down-sampled spectrum. Although filter bank analysis has engineering benefits such as reducing the number of spectral coefficients, it is primarily psycho-acoustically motivated with the goal of modifying the spectrum to more accurately represent the human perception of speech.

Discrete Cosine Transform

The filter bank energies give an adequate representation of the spectrum, but they are correlated with each other. A discrete cosine transform (DCT) is applied to convert the filter bank energies to the cepstral domain in which the cepstral coefficients are less correlated. Also, since Euclidean distance in the cepstral domain is equal to

$$\int \left(\log S_1(\omega) - \log S_2(\omega)\right) d\omega \,, \tag{2.7}$$

it maintains the distance relationships between spectra while providing a more computationally efficient distance measure. The discrete cosine transform is defined by

$$c_n = \sum_{k=1}^{N_f} \Theta[k] \cos\left[n(k-0.5)\frac{\pi}{N_f} \right], \tag{2.8}$$

where N_f is the number of filter bank energies, and $\Theta[k]$ is the sequence of filter bank energies.

Finally, because cepstral coefficients typically decrease sharply in magnitude as their index increases, cepstral liftering is sometimes performed to normalize their magnitudes (Deller *et al.*, 1993:378). A common liftering function is

$$c'_n = \left(1 + \frac{L}{2}\sin\frac{\pi n}{L}\right) c_n \,, \tag{2.9}$$

where L is a parameter usually defined to be n or slightly greater than n.

PLP Analysis

The PLP feature extraction model as described by Hermansky (1990) is shown in Figure 2.6. Each part of the block diagram is explained in the sections below.

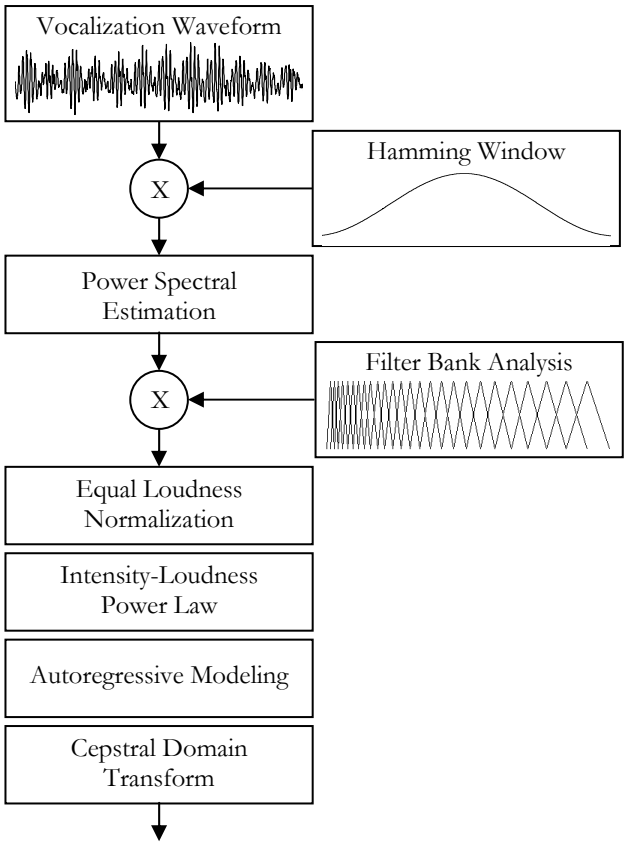


Figure 2.6 – PLP Block Diagram

Hamming Window / Power Spectral Analysis

The first component of the PLP feature extraction model is windowing the vocalization as described in the MFCC section above and calculating the power spectrum, $S(\omega)$, for each window. As in the MFCC model, many different power spectrum estimation techniques could be used, but the FFT is the most common. Refer to the corresponding MFCC section for more details.

Filter Bank Analysis

The filter bank analysis component in the PLP model is slightly different from MFCC filter bank analysis. First, the Bark scale (Schroeder, 1977) is used to logarithmically warp

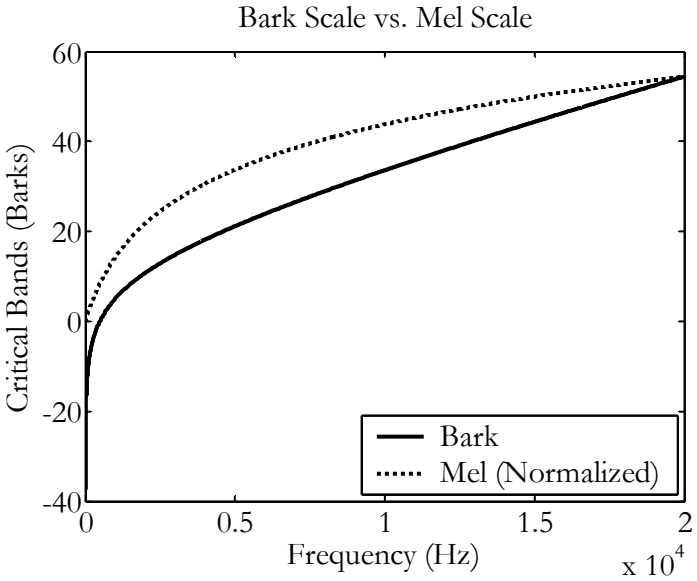


Figure 2.7 – Bark Scale Compared to Mel Scale

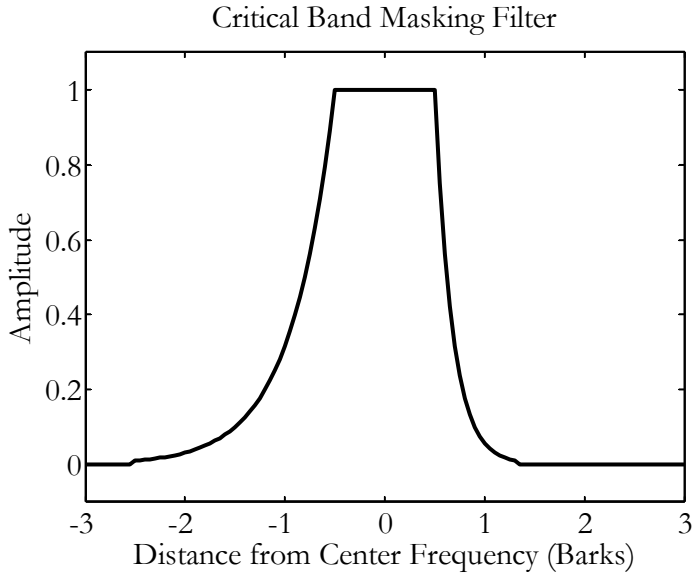


Figure 2.8 – Critical Band Masking Filter

the spectrum instead of the Mel scale. The Bark scale is based on critical bandwidth experiments and can be expressed as

$$f_{Bark} = 6\ln\left(\frac{f}{600} + \left(\left(\frac{f}{600}\right)^2 + 1\right)^{0.5}\right). \qquad 2.10$$

The Bark scale is shown in Figure 2.7 compared to the Mel scale.

The second difference is that perceptually shaped filters (Fletcher, 1940) are used instead of triangular filters. These perceptually shaped filters are more computationally expensive, but better approximate the human critical band masking filter shape. These filters can be expressed by

$$\Psi(f_{Bark}) = \begin{cases} 0 & f_{Bark} < -1.3 \\ 10^{2.5(f_{Bark}+0.5)} & -1.3 \le f_{Bark} \le -0.5 \\ 1 & -0.5 < f_{Bark} < 0.5 \\ 10^{-1.0(f_{Bark}-0.5)} & 0.5 \le f_{Bark} \le 2.5 \\ 0 & f_{Bark} > 2.5 \end{cases}, \quad 2.11$$

where f_{Bark} is the distance in Barks from the center frequency of the filter. The shape of one of these filters is shown graphically in Figure 2.8.

Equal Loudness Normalization

After the filter bank analysis, PLP performs a number of perceptual-based operations. The first of these is to apply an equal loudness curve to the filter bank energies to emphasize those frequencies to which humans are more sensitive and suppress the others. Hermansky (1990) used the equal loudness curve derived by Makhoul and Cosell (1976) and based on the human 40dB sensitivity curve determined by Robinson and Dadson (1956). This curve can be expressed by

$$E(f) = \left(\left(f^2 + 1.44\times10^6\right)f^4\right) \Big/ \left(\left(f^2 + 1.6\times10^5\right)^2\left(f^2 + 9.61\times10^6\right)\right). \quad 2.12$$

An alternative formulation which includes a term which takes into account humans' decreased sensitivity to higher frequencies is

$$E(f) = \left(\left(f^2 + 1.44\times10^6\right)f^4\right) \Big/ \left(\left(f^2 + 1.6\times10^5\right)^2\left(f^2 + 9.61\times10^6\right)\left(f^6 + 1.56\times10^{22}\right)\right). \quad 2.13$$

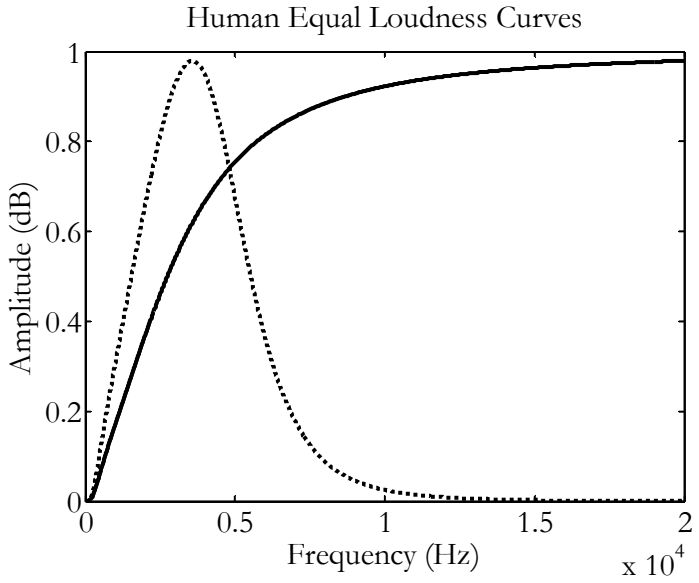


Figure 2.9 – Human Equal Loudness Curves

Both curves are shown in Figure 2.9. The equal loudness curve with the high frequency term (dotted line) has been normalized to the curve without the high frequency term (solid line) for clarity.

Intensity-Loudness Power Law

The next component in the PLP model applies the intensity-power law, which relates the power of the audio signal to perceived loudness. The relationship is defined as

$$\Phi[i] = \Xi[i]^{0.33}. \qquad 2.14$$

Stevens was the first to propose this law and performed experiments to validate his hypothesis (1957). This operation also compresses the power spectrum. As a result, the spectrum can be more accurately approximated by an all-pole autoregressive model of low order in the next step even though this was not the original motivation.

Autoregressive Modeling

The remaining components of the PLP model are concerned with transforming the perceptually modified filter bank energies into more mathematically robust features. An all-pole autoregressive model is used to approximate $\Phi(f_i)$ to smooth the spectrum and reduce

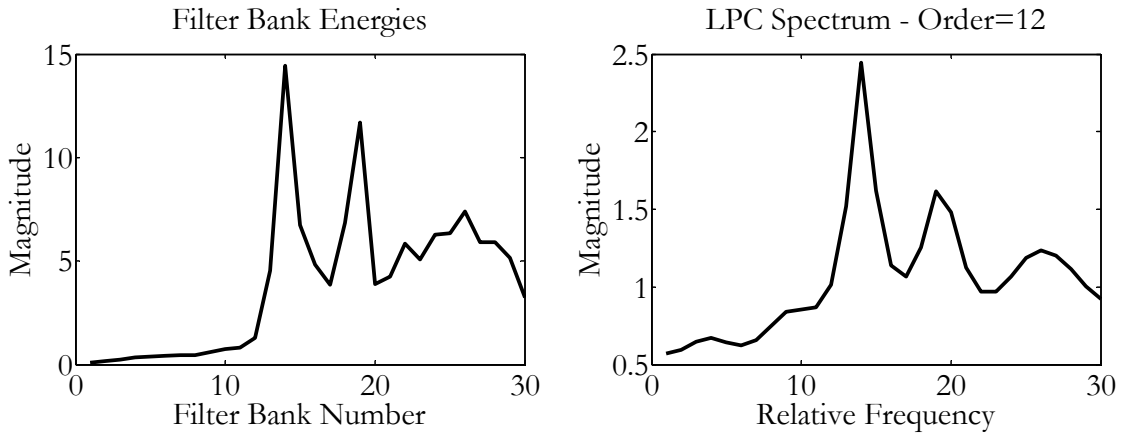


Figure 2.10 – Autoregressive Modeling Example

the number of coefficients. The model is derived using the Yule-Walker equations as specified in Makhoul (1975). Hermansky (1990), in the original PLP paper, determined that a fifth order model was adequate to capture the first two formants of human speech while suppressing the speaker specific details of the spectrum. An example of the effect of autoregressive modeling is shown in Figure 2.10. The plot on the left in the figure is a plot of the filter bank energies after perceptual modeling. The plot on the right is the smoothed filter bank energies using a 12th order autoregressive model. Notice how the LPC spectrum captures the relative heights of the three main peaks in the filter bank energies while smoothing the envelope. For a more detailed background on LPC analysis and autoregressive modeling, see Haykin (2002:136).

Cepstral Domain Transform

Finally, the autoregressive model coefficients are transformed to the cepstral domain. It has been shown that Euclidean distance is more consistent in the cepstral domain than when used to compare autoregressive coefficients (Deller *et al.*, 1993). There are more complicated distance metrics that are consistent for autoregressive coefficients, such as Itakura distance (Itakura, 1975), but these are much more computationally intense.

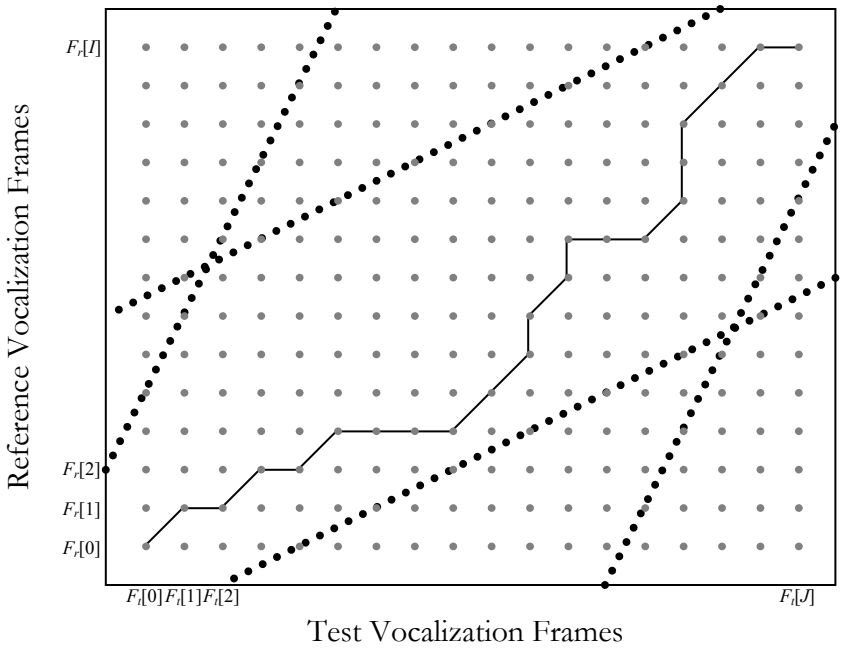


Figure 2.11 – Dynamic Time Warping

The autoregressive coefficients can be transformed to cepstral coefficients using the regressive equation

$$c_n = -a_n - \frac{1}{n}\sum_{i=1}^{n-1}(n-i)a_i\, c_{n-i}\,, \qquad 2.15$$

where a_n are the autoregressive coefficients. Liftering can also be done on these coefficients as in the MFCC model, but it is usually not necessary if the coefficients are modeled statistically because the smaller variance for the higher cepstral coefficients will account for the smaller range of the higher indexed coefficients.

Dynamic Time Warping

Dynamic time warping (DTW) is the most widely used template matching technique in speech recognition. Although DTW has been replaced in most state-of-the-art systems by stochastic models, its simplicity makes it desirable for small-scale systems. DTW has also been used for bioacoustic signal classification (Buck and Tyack, 1993). DTW is a non-linear

time alignment technique which compensates for temporal variation. It is most useful for isolated-word speech recognition classification tasks, but it can also be used for call-dependent speaker identification.

DTW is a dynamic programming technique (Silverman and Morgan, 1990), which compares a test vocalization to a reference vocalization frame-by-frame. This comparison can be best visualized by a grid as in Figure 2.11. The frames of the test vocalization, $F_t[J]$, are enumerated along the horizontal axis, while the frames of the reference vocalization, $F_r[I]$, are enumerated along the vertical axis. At each point in the grid, (i,j), the distance between reference vocalization frame i and test vocalization frame j is calculated. The least cost path from (0,0) to (I,J) is then determined via dynamic programming. The dynamic programming algorithm is guided by both global and local path constraints. Global path constraints, shown by the thick, dotted lines in the figure, assures that the path does not cause an unrealistic warping of the test vocalization. Without the global path constraints, the path could match test frames which occur near the end of the vocalization with reference frames that occur early in the vocalization if similar phonemes are present in both places or if silence occurs before and after the utterances. This matching of ending frames with beginning frames would be highly unlikely.

Local path constraints define the possible transitions between grid points. In this example, the path can either go one point vertically, one point horizontally or one in each direction to create a diagonal path. Each of these local paths is usually weighted to make the diagonals slightly more costly. For example, while the horizontal and vertical paths may be weighted at unity, the diagonal path might be weighted at 2 since the one diagonal path is the summation of one horizontal and one vertical movement. The total cost of the path can be

used as a similarity measure and the path through the grid shows how the frames from the test vocalization and template match.

To train a DTW system, templates of each vocalization type must be created. Although one vocalization from each class can be picked as the template, it is advantageous to use a number of training vocalizations to create more generalized templates. To train a template based on multiple vocalizations, all training examples are time-warped to the median-length example. The path from the time-warping is used to match up frames from all training examples, and the mean and variance of each template frame is determined from all matching frames. More information about DTW can be found in Deller *et al.* (1993).

Hidden Markov Models

The hidden Markov model (HMM) is the most popular model used in human speech processing to model the different segments of a speech waveform (Juang, 1984; Rabiner, 1989; Rabiner and Juang, 1993). Nearly all modern, state-of-the-art speech processing systems are based on some derivative of the classic HMM. An HMM consists of a number of states which are connected by transition arcs, and can be thought of as a statistically based state machine. An HMM is completely defined by its transition matrix, A, which contains the probability of the system transitioning from state i to state j, and state observation probabilities, $b_i(o)$, which model the observed parameters of the system while in that state. Because the transition matrix is two-dimensional, the system is assumed to hold to the Markov property: the next state is dependent only on the current state as opposed to states it may have been in previously. Gaussian Mixture Models (GMMs) are commonly used to model the state observation probability densities.

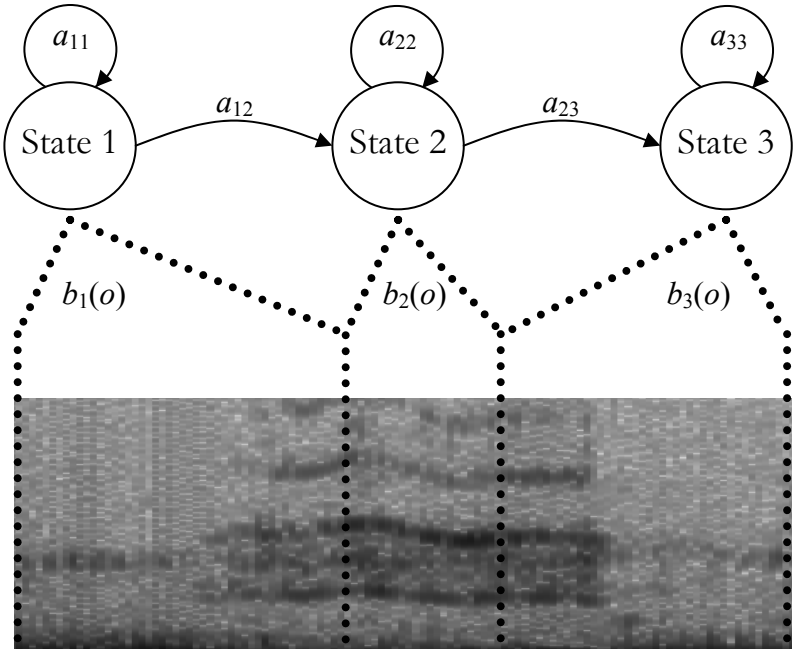


Figure 2.12 – Hidden Markov Model

In application to time series, each state of the HMM represents a quasi-stationary portion of the time signal, and each complete HMM model represents a language structure such as a phoneme, word, or sentence. Most large vocabulary speech systems use phoneme-based models, while smaller vocabulary systems, such as numeric digit recognition systems, use word-based models to more efficiently use training data. Usually, the HMM is constrained to only allow transitions from left to right by one state at a time to model the time-dependent nature of speech systems. A typical left-to-right HMM is shown in Figure 2.12. In this example, the HMM is modeling an African elephant trumpet. As shown in the figure, the three states map to the beginning, middle, and end of the vocalization.

An HMM can be trained to model sequences of features extracted from a particular set of vocalizations using a variety of algorithms. The most popular is the Baum-Welch reestimation method (Baum, 1972; Baum *et al.*, 1970). The Baum-Welch method is an expectation maximization algorithm (Moon, 1996) which maximizes the output likelihood of

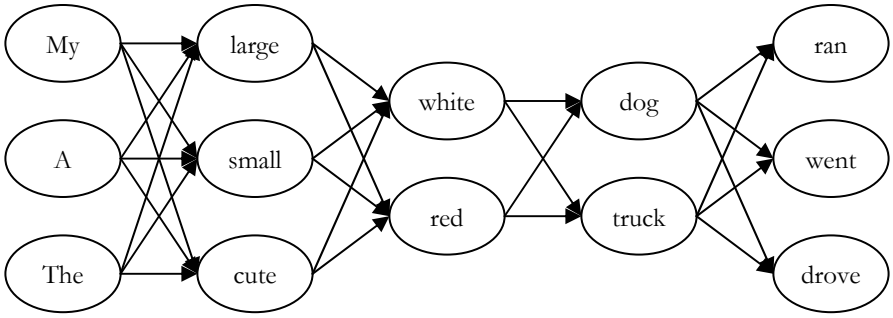


Figure 2.13 – Word Network

the model with respect to the training data over all possible state transitions. Baum-Welch is popular because of its quick convergence properties; it normally converges after 4-5 iterations over the training dataset. It is also guaranteed to improve the output likelihood of the training dataset with each iteration.

The Viterbi algorithm is used to match an HMM to a sequence of features from a new vocalization (Forney, 1973). The Viterbi algorithm, a dynamic programming algorithm, determines the probability of a vocalization fitting the HMM over the most likely state sequence. The typical use of the Viterbi algorithm in a speech processing application is to find the maximum likelihood path through a network of connected HMMs. A very simple network of word-based HMMs is shown in Figure 2.13. The network is defined by a grammar which can either be fixed or statistical, in which each transition between models is given a probability that is incorporated into the Viterbi algorithm. For instance, in the above example, it is likely that the node "dog" would have a higher transition probability to the node "ran" than it would to the node "drove".

The statistical basis of the HMM is the biggest benefit over template-based models since other probabilistic models can be incorporated directly into the training and recognition process. Probabilistic language, duration, or grammar models are examples of models that can be incorporated directly.

Gaussian Mixture Models

Gaussian Mixture Models (GMMs) are parametric probability density functions commonly used in statistical classification systems (Deller *et al.*, 1993:706). The probability density function is represented by a weighted sum of Gaussian distributions,

$$p = \sum_{m=1}^{M} w_m N(\mu_m, \sigma_m), \qquad 2.16$$

where M is the number of mixtures, and w_m is the mixture weight. If the mixture weights sum to 1, the GMM gives a valid distribution. Given an adequate number of mixtures, a GMM can model any arbitrary distribution (Deller *et al.*, 1993:706). GMMs are used in speech processing for speaker identification models and as the probability distribution model for HMM state observation probabilities.

Language Models

Current speech systems rely heavily on language models to achieve their high recognition accuracies (Allen, 1995; Harper *et al.*, 1999; Johnson *et al.*, 1998). Studies have shown that acoustics alone can currently achieve phoneme recognition accuracies of around 65% on clean speech (Chengalvarayan and Deng, 1997). With the addition of noise or mismatched training and testing conditions, this accuracy drops significantly. Language models aid in the recognition process by making sure that the recognized phoneme or word sequences make linguistic sense.

Language models can be grouped into two general types, fixed and statistical. Fixed grammars require the recognized text to conform to a specific syntax. Fixed grammars are useful when the scope of the conversation is constrained, as for making airline reservations or stock trading. In these cases, the user can be prompted to use a specific syntax such as "I would like to fly to <destination> from <origin> on <date>". In this case, only the

destination, origin, and date are variable, and the recognition system can use surrounding words to aid in the recognition process.

Statistical language models are used more often in unconstrained domains such as closed captioning for movies or television, or dictation applications. Statistical language models use training text from the domain to assign probabilities to groups of words that commonly appear together. For instance, "fire truck" would probably have a higher language probability than "blanket truck." Given a statistical classification model, these probabilities can be applied directly to the recognition process. The longer word sequences that are modeled, the larger the language model. Trigram language models, where groupings of three words are modeled, are common, but 4-grams and 5-grams are used on occasion when there is a large amount of training data.

Summary

The research presented in this dissertation borrows heavily from previous work in speech processing, specifically speech recognition systems. Feature extraction models and classification models common to speech processing discussed in this chapter are modified to apply them to animal vocalization classification tasks. The next chapter will provide background in current bioacoustic signal analysis techniques and discuss some current research in the classification of bioacoustic signals.

Chapter 3

BIOACOUSTICS

Background

The field of bioacoustics, the study of animal vocalizations, has received increased attention in recent years with the advent of new recording and analysis technologies. The main goal of bioacoustics is to determine the role of animal vocalizations in the communication process. Improved, less invasive recording technologies have allowed researchers to collect better data in the animal's natural habitat, and easier to use analysis tools have allowed researchers to study the data without an extensive knowledge in signal processing. The field of bioacoustics is multi-disciplinary, with biologists, animal behaviorists, neurologists, psychologists, and more recently, engineers contributing to the field.

Although analysis techniques have improved greatly in recent years, there is still a massive technology gap between animal and human vocalization analysis techniques. One reason for this gap is the lack of knowledge about how some species produce and perceive sound. Another reason is the lack of interest from the speech community to adapt techniques to animal vocalization analysis. Analysis tools that perform vocalization classification, speaker identification, word spotting, and behavior correlation could be useful to the bioacoustics community.

Current Features

Although the speech processing community has a few sets of standardized features, the bioacoustics community does not. Usually, features are measured by hand from spectrograms for each vocalization. The features are usually vocalization-based, with one

value for the entire vocalization, as opposed to the frame-based features popular in speech processing discussed in chapter 2. Although the set of features is not standardized, there are a number of features commonly used including duration, maximum, minimum and average fundamental frequency, average amplitude, and number of syllables in the vocalization. Features that are calculated once for each vocalization work well with statistical test techniques such as the t-test, Chi-Squared test, MANOVA and factor analysis. This is one main reason why traditional bioacoustic features have been measured once for each vocalization.

The most complete effort to standardize features used in the bioacoustic field is the AcouStat project. AcouStat, created by Fristrup and Watkins (1992; 1994) and later modified by DiMarzio and Watkins (Watkins *et al.*, 1998), automatically extracts 120 different features from both the time and frequency domains. As is typical in the bioacoustic field, features are calculated over the entire vocalization. Some features that AcouStat calculates include average amplitude, duration, and the average peak frequency. AcouStat could be used for various species, but it is not very popular in the bioacoustics field.

In typical bioacoustic studies, the features are usually put through a dimension reduction mechanism, such as principle component analysis (PCA) or discriminant function analysis, to select the most important features (Fristrup and Watkins, 1992; Leong *et al.*, 2002; Owren *et al.*, 1997; Recchia, 1994; Riede and Zuberbühler, 2003; Sjare and Smith, 1986b). The reduced-dimension data can be plotted for visualization and classification.

There has been recent work using features extracted multiple times for each vocalization using temporal windows (Buck and Tyack, 1993; Murray *et al.*, 1998; Schön *et al.*, 2001). The trouble with using these features is that they do not work well with traditional statistical tests designed for one measurement for each example. Therefore, most of the studies using

frame-based features apply other classification methods. Artificial neural networks (ANNs) are the most common. The studies which use more advanced classification methods will be discussed in their appropriate section below.

The results of these studies show that frame-based features can be used to classify bioacoustic signals with high accuracy. Many different classification tasks have been implemented for diverse species using various classification models. Some of the more advanced systems are discussed in the next section, grouped by the classification task.

Classification Tasks

Repertoire Determination

One common task in bioacoustics is the determination of a species' repertoire of vocalizations (Berg, 1983; Cleveland and Snowdon, 1982; Sjare and Smith, 1986b). Normally, this is achieved by analyzing spectrograms of the various vocalizations and then grouping similar sounds into a single call type. Sounds are broken into types based on harmonic structure, pitch contour, whether the vocalization is pulsed, or other criteria. A pulsed call is a vocalization in which the spectral energy is not continuous throughout the vocalization. Instead, the vocalization is rapidly modulated between "on" states where the animal is actively calling and "off" states where there is no spectral energy. Whenever possible, behavior recorded in conjunction with the vocalization is used to help distinguish between the different types of sounds. Once the basic sound types are identified, a language structure can be hypothesized for those species whose vocalizations consist of a number of different syllables, such as bird or whale song.

Sometimes, classification or statistical analysis techniques are used to validate the difference between sounds in the repertoire. Murray *et al.* (1998) used two features, peak frequency, and duty cycle, to categorize the repertoire of false killer whales. Duty cycle is the

percentage of the time the waveform amplitude is greater than the average amplitude value. The two measurements were made every 11.6ms using 11.6ms (512 points at 44.1kHz) windows. To keep the number of measurements for each vocalization the same, only the first 30 measurements of each feature for each vocalization were used as input into an unsupervised ANN. The ANN was able to distinguish between the two main types of vocalizations, ascending whistles and low-frequency pulse trains.

Ghosh *et al.* (1992) experimented with various types of ANNs and statistical classifiers to perform call type classification. Wavelet coefficients calculated over the entire signal and the duration of the signal were used as the features for each classifier. All proposed classifiers had comparable performance with a correct classification rate of about 98% using relatively clean data.

Riede and Zuberbühler (2003) analyzed the difference between the Diana monkey's leopard and crowned eagle alarm vocalizations, both of which are pulsed. LPC analysis was performed on the two vocalization types to identify the formants in each vocalization. Both the pulse duration and the location of the first formant at the beginning of the vocalization were different between the two types of vocalizations. However, the first formant at the end of the vocalizations is not different. Therefore, the leopard alarm call has a larger downward movement of the first formant from the beginning to the end of the vocalization. This consistent difference between the calls supported the hypothesis that the alarm calls are distinct vocalizations in the repertoire.

Species Determination

Automatic classification systems which can determine the species that made a vocalization have been developed recently. Chesmore (2001) created a system and then demonstrated its success on a number of insect species (Orthoptera) from Great Britain and

separately on a number of avian species from Japan. The system used time-domain based features based on the number of minima and maxima in the waveform during one pitch cycle and the length of the pitch cycle. These features essentially capture the shape of the waveform. These features are then used for input into a feed-forward ANN. The system was designed to be easily implemented in basic integrated circuits, which is the main reason for the lack of spectral analysis.

Anderson (1999) and Kogan and Margoliash (1997) have both compared the performance of HMM and DTW systems in the determination of species from bird song. Human speech features were used to parameterize the vocalizations in both studies. The DTW systems performed better when training data was limited, but the HMM systems were more robust to noise. HMMs also did better when classifying vocalizations that varied from the usual song types. These results are typical of those reported in speech processing literature on the differences between DTW and HMMs.

Individual Identification

Recently, there have been many articles published on the identification of the individual making the vocalization. Systems that perform speaker identification are highly desirable because the determination of speakers while recording data in the animals' natural habitat can be extremely challenging when the speaker is hidden from view. There is evidence that parents from a number of species can identify their young from their vocalizations (Charrier *et al.*, 2002; Insley, 2000; Insley *et al.*, 2003). This knowledge has led to the exploration of whether a system can be constructed to identify the vocalizing individual from a population.

There are two main techniques used to show the individuality of the vocalizations. The first involves playback experiments which are most commonly done in the parent-offspring identification experiments (Charrier *et al.*, 2002; Goldman *et al.*, 1995; Insley, 2000, 2001;

Insley *et al.*, 2003). The young's vocalizations are recorded and played back to the parent. If the parent's response to its own young's vocalizations is different from the response to unknown vocalizations, then it is concluded that the parent can recognize its young's own call. In one these studies (Charrier *et al.*, 2002), the vocalizations were modified to determine the aspect of the call that the parent uses to determine whether it is its young. Insley *et al.* (2003) showed that the young can also identify their parents.

The other technique used to determine whether speaker identification is possible with animal vocalizations involves extracting features from the vocalization and applying a statistical testing technique to determine whether the features extracted from one individual are different from those features extracted from the other individuals (Darden *et al.*, 2003; Durbin, 1998; Goldman *et al.*, 1995; Insley, 1992; Sousa-Lima *et al.*, 2002). All of these studies use features measured from the spectrogram, spectrum, or waveform. The features are calculated on a vocalization basis for use in statistical tests as is typical in bioacoustics research. The statistical test varied for each study based on the type of data and experimental setup, but a statistical difference in the features collected from each individual verified that the individuals can be determined from acoustic features of the vocalization. In some of the studies, multivariate analysis was performed to determine the most significant features in separating the individuals. Charrier *et al.* (2003) used statistical tests on extracted frequency features to show that fur seal vocalizations change with age and playback experiments to show that female fur seals remember the vocalizations of their young even after the young are grown.

Some studies have used a system more closely related to typical speech processing systems for speaker identification. In Campbell (2002), a feed-forward ANN was trained to identify the speaker out of a population of 26 female Steller sea lions. The frequency

spectrum, averaged over the vocalization, was used as input into a back-propagation trained neural network with 26 outputs, one for each subject. The study had a classification accuracy of up to 70.69% on the testing dataset. Buck and Tyack (1993) used Dynamic Time Warping (DTW) to time-align pitch contours to identify individual Bottlenose Dolphin (*tursiops truncatus*) whistles. The method correctly classified 29 of 30 whistles from five different dolphins. The total cost of the DTW path was used as a similarity measure to provide a measure of confidence of the match between two whistles. See chapter 2 for a more complete description of the DTW classification model.

Stress Detection

With the passing of more restrictive animal welfare laws, the detection of stress, especially in domesticated animals, has become an important issue. One suggested method of detecting stress has been through monitoring the vocalizations of the animals. Schön *et al.* (2001) outlines one of the most complete systems for detecting stress in domestic pigs. Linear Predictive Coding (LPC) coefficients, which are sometimes used to derive cepstral coefficients in speech analysis, were used to quantify the vocalizations. Twelve LPC coefficients derived using 46.44ms (1024 point at 22kHz) windows were used as input into an unsupervised ANN. Screams, stress vocalizations, were correctly classified with greater than 99% accuracy, while grunts, non-stress vocalizations, were correctly classified greater than 97.5% of the time.

Call Detection

The ability to detect bioacoustic signals in background noise would drastically speed up the transcription and segmentation of collected data. A call detection system could be used to prevent unnecessary human-animal interaction by redirecting ships around clusters of animals when their vocalizations are detected. Potter *et al.* (1994) used individual pixels

(time-frequency bins) from a smoothed spectrogram and an ANN to detect bowhead whale song endnotes in ocean background noise. Low-resolution spectrograms were calculated for each vocalization using a Δf of 63.5Hz and Δt of 128ms to generate an 11 x 21 spectrogram. The supervised ANN classification of bowhead whale endnotes was more accurate than a system using spectrogram cross-correlation with higher resolution spectrograms to classify the sounds.

Weisburn *et al.* (1993) compared the performance of a matched filter and an HMM system for detecting bowhead whale (*Balaena mysticetus*) notes. While the matched filter used a spectrogram template, the HMM experiment used the three top peaks in the spectrum as features for an 18-stage model. Although the HMM detected 97% of the notes compared to 84% for the matched filter, the HMM also had 2% more false positive detections.

Niezrecki *et al.* (2003) compared a number of methods for the detection of West Indian manatee (*trichechus manatus latirostris*) vocalizations. A simple spectral peak threshold method, a harmonic threshold method, and an autocorrelation method based on the energy in four frequency bands were the three methods compared. While the autocorrelation method yielded the best detection accuracy at 96.34%, the harmonic threshold method had the fewest false positives at 6.16%. Considering that the methods are based on simple thresholds, they performed extremely well.

One of the more popular software systems for detecting calls is Ishmael, developed by Mellinger (2002). Ishmael, along with displaying spectrograms, can find similar calls in a recording using spectrogram correlation (Mellinger and Clark, 2000). Spectrogram correlation compares the similarity between two spectrograms. Therefore, by defining a template spectrogram for a vocalization, similar vocalizations can be found through

comparison. The software has been used in a number of studies including Mellinger *et al.* (2004) where it was used to detect right whale (*Eubalaena japonica*) calls.

Summary

Bioacoustics research has made recent strides in the analysis of bioacoustic signals but has yet to standardize on a single feature extraction model or classification model. Although classification systems have been built for a variety of tasks, they are customized to the task and species under study. The next chapter will discuss a standardized methodology and framework for analyzing animal vocalizations which is adaptable to different species and tasks.

Chapter 4

METHODOLOGY

Background

To successfully analyze animal vocalizations, the feature extraction and classification models need to reflect the perceptual abilities of the animal species under study. Animals' sensitivity to various frequencies is different than humans and they lack a formal language made up of phonemes, words and sentences, therefore it is clear that human speech processing techniques need to be modified for each species under study. These modifications to the feature extraction and classification models will be presented as the gPLP framework.

This chapter outlines the various signal processing and classification model changes made during the course of this research and how the gPLP framework developed out of these changes. Examples of the effect of the gPLP feature extraction model on the spectrum are displayed and analyzed. Also, a method for applying gPLP coefficients to traditional bioacoustics statistical tests is presented. Finally, the gPLP framework is applied to two species, African elephants and beluga whales, and the results presented in the following two chapters.

Classification Models

The two different classification models used in the gPLP framework are dynamic time warping (DTW) and hidden Markov models (HMM). As discussed in chapter 2, DTW is a template-based model while the HMM is a statistical model. DTW was once popular in speech processing but has since been replaced in large by HMMs due to the statistical nature and improved robustness to noise of the HMM. In the following sections, the parameters

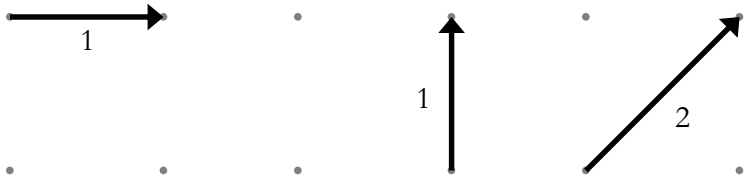


Figure 4.1 – Valid DTW Local Paths and Weights

used in both models will be discussed in reference to the typical parameters in speech processing.

Dynamic Time Warping

A dynamic time warping classification system was written in MATLAB® for the gPLP framework. The system includes the traditional training algorithm and dynamically programmed recognition algorithm as discussed in chapter 2. After a number of trial runs, the paths being generated by the algorithm were examined and determined to be realistic. Therefore, global path constraints were not implemented in either the training or testing algorithm. The three valid local paths and their weightings used in the DTW system are shown in Figure 4.1. This set of valid local paths was originally considered by Sakoe and Chiba (1978). Euclidean distance between the test and reference feature vectors was used as the distance metric.

Hidden Markov Model

The Hidden Markov Toolkit (HTK) version 3.2.1 from Cambridge University's Engineering Department (2002) was used to implement the hidden Markov models (HMMs). This package was chosen because of its flexibility and the inclusion of various types of language models. It is also open source, therefore the new feature extractions models discussed in later sections could be included. A number of parameters were varied to find optimal values for the parameters including the number of states in the HMMs and the inclusion of a silence model. These variations are discussed in the results chapters.

Feature Extraction Models

Two different feature extraction models were used during the development of the gPLP framework, Mel frequency cepstral coefficients (MFCCs) and generalized perceptual linear predication (gPLP). While the MFCC feature extraction model is more widely used in human speech processing, the gPLP feature extraction model is a new, novel model which borrows heavily from the perceptual linear prediction (PLP) model developed by Hermansky (1990). The gPLP model can incorporate perceptual information from the species under study such as range of hearing, sensitivity to different frequencies, and discrimination between closely spaced frequencies.

After discussing initial efforts to improve the MFCC feature extraction model, the gPLP model will be outlined along with methods for constructing alternative warping scales and equal loudness curves from commonly available experimental data. Finally, examples are provided which visualize the effect of the gPLP feature extraction model on the spectrum.

Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCCs) were the initial features used as input to the classification models. At first, standard MFCC model parameters employed in speech processing were used to provide a baseline for future improvements. However, as the classification experiments were performed, it became clear that these features did not capture some of the very low frequency characteristics well (less than 100Hz) which are prominent in many species, including African elephants. Therefore, the following modifications were made to the MFCC feature extraction model to capture these very low frequency characteristics better.

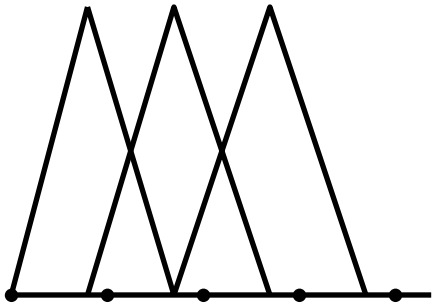


Figure 4.2 – Filter Bank Range Compression

Filter Bank Range Compression

As discussed in chapter 2, the MFCC feature extraction model involves the computation of filter bank energies. The placement of the filters in the filter bank has a large effect on the value of the resulting cepstral coefficients. This large effect is demonstrated by the improved classification accuracies in speech recognition experiments when the filters are spaced according to the Mel frequency scale instead of linearly (Davis and Mermelstein, 1980).

In a typical speech processing system, the filters are spaced across the length of the spectrum, from 0Hz to the Nyquist frequency. However, in the case of very low frequency vocalizations, this places a large number of filters above the energy range of the vocalization. The filters placed above the energy range of the vocalization do not contribute to the accuracy of the extracted features, but instead add noise to the calculation of the features. However, by placing all of the filters within the known energy range of the vocalizations, the problem of filters contributing noisy information is addressed. For example, if a set of vocalizations are known to be in the 10Hz – 300Hz range, then the filters can be spaced according to the Mel scale between those frequencies to focus on that frequency range.

Fourier Transform Padding

Although compressing the frequency range of the filters was effective in increasing classification accuracies, it presented another problem. As the distance between the center frequencies of the filters decreased due to the filter bank range compression, the filters contained fewer points of the spectrum. This effect can be seen in Figure 4.2. Notice that the leftmost filter has only one point in the spectrum contributing to the filter energy calculation. The other two filters have two spectral points contributing to the filter energy calculation; however, one point in each filter only contributes a small amount since it is near the edge of the filter. The lack of points contributing to the energy calculation makes it inaccurate.

One way to compensate for this lack of spectral points is to zero pad the signal before the Fourier transform to interpolate between the existing points. It is important to note that padding the signal does not actually increase the precision of the spectrum, but instead smoothes and interpolates between the fixed precision points which are spaced at $1 / w_s$ Hz, where w_s is the window size in seconds.

The effect of this interpolation is to create more points to contribute to the calculation of the filter energies. Consequently, the filter energies are much more accurate. This leads to more accurate MFCCs and consequently, to more stable classification accuracies when the parameters are perturbed to a small degree. As the results chapters will show, an interpolation of more than 4 times the original spectral resolution did not lead to more stable classification accuracies, inferring that after a certain number of points are used to calculate the filter energies, the improvement in accuracy is small.

GPLP COEFFICIENTS

As these changes to the MFCC feature extraction model were being explored, it became evident that if the feature extraction model could be tailored to the perceptual ability of the

species under study, the classification accuracies could be improved. Perceptual linear prediction (PLP) analysis (Hermansky, 1990), based on human perception, is a good starting point for constructing a generalized PLP (gPLP) feature extraction model, which replaces human perceptual information with information available on specific species.

Although PLP is based on the source-filter model and was originally designed to suppress excitation information while accenting the filter characteristics, the use of higher order autoregressive modeling in the feature extraction model can capture excitation and harmonic information. Vocal tract features carry the majority of the information content in human speech, but traditional bioacoustic features tend to concentrate more on the excitation characteristics of the vocalization because animal vocalizations tend have less dynamic spectral envelopes, but can have many more harmonics. gPLP can model both harmonically rich sounds as well as vocalizations with a complex filter structure by adjusting parameters of the feature extraction model.

The block diagram of gPLP analysis is in Figure 4.3. Similar to the PLP block diagram, the gPLP block diagram adds an additional step and includes experimental tests that can be used to construct the various species-dependent aspects of the model. Each stage of the block diagram is discussed in the sections below along with the adaptations required to apply the model to a particular species.

Pre-Emphasis Filter

The first component of the gPLP feature extraction model is the pre-emphasis filter. The purpose of the pre-emphasis filter is to normalize the spectral tilt that results from the general nature of the vocal tract filter. Although this phenomenon was first described in human speech spectra, it is common in the vocalizations of other species as well. To normalize for spectral tilt, the higher frequency components of the signal are emphasized to

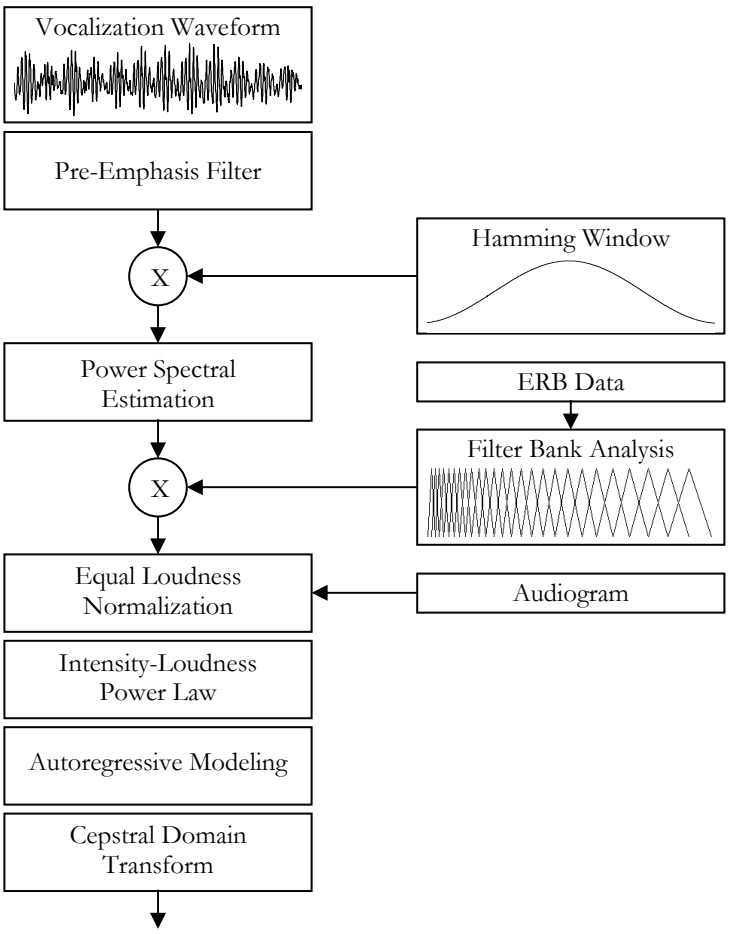


Figure 4.3 – gPLP Block Diagram

make their magnitudes more comparable to the lower frequency spectral values. If this pre-emphasis is not performed, the lower frequency formants dominate during the calculation of the cepstral coefficients and the higher frequency formants are largely ignored because they have a lower dynamic range. Although not part of the PLP model as described by Hermansky (1990), experimental results show that the addition of a pre-emphasis filter improves the robustness of the extracted features.

To perform the pre-emphasis, the digitized vocalization waveform, $s[n]$, is modified by a high-pass filter of the form

$$s'[n] = s[n] - \alpha s[n-1], \tag{4.1}$$

where α is typically near 0.97. A value of α=0.0 creates an all-pass filter, while a value of α=1.0 creates a high-pass filter with linear magnitude response where the normalized magnitude is 0 at 0 Hz and 1 at the Nyquist frequency. The value for α can be increased to further emphasize higher frequencies or decreased to make the filter closer to an all-pass filter. Experimental results are used to determine the best value of α for each particular species. A value of α=0.97 is a good beginning estimate for the best value of α for a species. Experimental results show that when other feature extraction parameters are optimized, the value of the pre-emphasis coefficient has little effect on the classification accuracy.

Hamming Window

The second component of the gPLP feature extraction model is the division of the vocalization into frames in preparation for spectral analysis. As discussed previously, the vocalization is framed to construct quasi-stationary frames for accurate spectral estimation. A windowing function is applied to each frame to reduce artifacts that would arise from performing spectral analysis on a non-windowed frame. See Oppenheim and Schafer (1999:465) for more information on the effects of windowing.

A Hamming window of the form

$$w[n] = 0.54 + 0.46\cos\left[2\pi n / (N-1)\right], \tag{4.2}$$

where N is the length of the frame, applied to each frame of the vocalization. The Hamming window is the most popular windowing function in speech analysis, but other windows may be used as long as their effects on spectral estimation methods are well understood. See Oppenheim and Schafer (1999:468) for a discussion on different analysis windows.

The frame size and frame step are important considerations for the feature extraction model. The frame size should be chosen to create frames with a sufficient number of pitch

periods to get a good estimate of the periodicity of the signal. Five is a typical number to include in speech processing applications. There is a trade off to the size of the frame. As the frame size increases, spectral resolution also increases because the frequency resolution, Δf, of the spectrum is related to the size of the window by

$$\Delta f = 1/w_s \,, \tag{4.3}$$

where w_s is the size of the analysis window in seconds.

However, as the frame size gets larger, the signal typically becomes less stationary over the frame. If the signal is not stationary across the entire frame to an adequate degree, the accuracy of spectral estimation methods declines significantly because the fine details of the spectrum will be averaged out. In general, the frame size should be as large as possible while ensuring stationarity across the frame. The degree of stationarity can often be approximated by examining the waveform and looking for consistency in the shape of the waveform between pitch peaks.

Overlaps of 1/3, 1/2, and 2/3 are common for the frame step size. By using a step size independent of the window size, both frequency and temporal resolution can be controlled. Frequency resolution can be determined by the window size, while temporal resolution can be determined by the step size. There is a trade-off however, of using too much overlap to create finer temporal resolution. Large frame overlaps lead to duplication of data because the spectrum from frame to frame will be very similar. On the other hand, a small frame overlaps may not sufficiently capture the dynamics of the signal. Signals with quickly changing characteristics should be analyzed with more overlap, while slowly changing signals can be analyzed with less overlap without losing information about the signal dynamics.

Power Spectrum Estimation

Once the signal is broken into frames, the power spectrum of each frame needs to be estimated. The Fast Fourier Transform (FFT) is the most common method for performing the estimation, but other power spectral estimation techniques could be used such as the MUSIC or Yule-Walker methods. A discussion of these various methods can be found in Stoica and Moses (1997). This study will use the FFT due to its popularity in the field of speech processing. The power spectrum, $P(\omega)$, can be calculated from the Fourier transform of the *w*th frame using the equation

$$P(\omega) = abs\left(FFT\left(s_w[n]\right)\right)^2. \qquad 4.4$$

When the signal is sampled at a low sampling rate, the frame can be zero padded before the FFT to increase the frequency resolution by interpolation. The effects of zero padding the signal are discussed previously in chapter 4.

Filter Bank Analysis

The next step in the gPLP feature extraction model is to apply a filter bank to the estimated power spectrum. The purpose of the filter bank is to model how the animal perceives the frequency spectrum. Filter bank analysis takes into account frequency masking through the filter shapes and the logarithmic cochlear position to frequency sensitivity map through the frequency warping function. After the filter bank is constructed, the energy in each filter is calculated. The sequence of filter bank energies represent a down-sampled and smoothed power spectrum that more closely approximates the response along the length of the cochlea.

Greenwood (1961) showed that many animal species perceive frequency on a logarithmic scale along the length of the cochlea by analyzing experimentally acquired frequency-position data. This phenomenon was described with the equation

$$f = A\left(10^{ax} - k\right), \tag{4.5}$$

where f is the frequency in Hertz, x is the distance from the stapes on the basilar membrane, and A, a, and k are constants defined for each species' physiology. Replacing x with perceived linear frequency, f_p, gives the frequency warping functions, which convert between real frequency and perceived frequency:

$$F_p^{-1}\left(f_p\right) = A\left(10^{af_p} - k\right), \text{ and} \tag{4.6}$$

$$F_p\left(f\right) = \left(1/a\right)\log_{10}\left(f/A + k\right). \tag{4.7}$$

The frequency scale is typically measured in Hz. However, since the scale in perceptual frequency is different, the unit of measure for the perceptual frequency scale is defined as pHz, perceptual Hertz. The Mel scale, discussed in chapter 2, is a specific implementation of these warping functions using the constant values of A=700, a=1/2595, and k=1. Greenwood (1990) calculated the constant values for a number of species by fitting equation 4.5 to frequency-position data.

If frequency-position data is not available, equal rectangular bandwidth (ERB) data can be used to derive the Greenwood warping constants using a method first developed for human auditory data (Zwicker and Terhardt, 1980). If the ERB data is fit by an equation of the form

$$ERB = \alpha\left(\beta f + \delta\right), \tag{4.8}$$

the Greenwood warping constants can be calculated using the following set of equations:

$$A = 1/\beta\,, \tag{4.9}$$

$$a = \alpha\beta\log_{10}\left(e\right), \text{ and} \tag{4.10}$$

$$k = \delta\,, \tag{4.11}$$

where e is Euler's constant, the natural logarithm base. The derivation of these equations is included in Appendix A.

The Greenwood warping constants can also be calculated using the hearing range of the species ($f_{min} - f_{max}$) and the assumption that k=0.88 in mammals. LePage (2003) found that most mammals had a value near k=0.88 when calculated from frequency-position data. LePage (2003) determined that this value is optimal with respect to tradeoffs between high frequency resolution, loss of low frequency resolution, maximization of map uniformity, and map smoothness. Non-mammalian species were not included in this study. Therefore, this assumption may not hold for those species. In non-mammalian cases, the aforementioned ERB method for deriving the Greenwood warping constants would be more appropriate. Using k=0.88, and the constraints that $F_p(f_{min})$=0 and $F_p(f_{max})$=1, the following set of equations can be used to find the other Greenwood warping constants:

$$A = \frac{f_{min}}{1-k}, \text{ and} \tag{4.12}$$

$$a = \log_{10}\left(\frac{f_{max}}{A} + k\right). \tag{4.13}$$

The derivation for these equations is in Appendix B.

If this method is used to derive the Greenwood constants, then the lowest filter in the filter bank must not extend below f_{min} because the real perceptual frequency is negative for real frequency values less than f_{min}. Negative values of f_p would cause problems with the calculation of the location of the filters in the filter bank.

Once the Greenwood warping function is derived for the species, the center frequencies of the filters in the filter bank are spaced linearly on the f_p axis. It is common for the filter bank to span the entire spectrum, from 0 Hz to the Nyquist frequency, $f_s/2$, where f_s is the

sampling rate. If this is the case, the distance between the center frequencies in perceptual frequency, Δcf_p, of the filters is given by

$$\Delta cf_p = \frac{F_p\left(f_{Nyquist}\right)}{n_f + 1}, \tag{4.14}$$

where n_f is the number of filters in the filter bank.

The number of filters to use is an important consideration. Hermansky (1990) suggests spacing the center frequency of the filters about one critical bandwidth apart, which is linear spacing in f_p units. If the ERB integral method is used to derive the Greenwood constants, the perceptual frequency, f_p, is already scaled one-to-one with the critical bandwidths. This means that the distance between f_p=2 pHz and f_p=3 pHz is exactly one critical bandwidth, and the filters can be spaced 1 pHz apart. However, if other methods are used to derive the Greenwood constants, f_p will not be scaled appropriately, and ERB data must be used to determine the number of filters to use to space the filters approximately one critical band apart. Humans have approximately 28 critical bands in their hearing range (20Hz – 20,000Hz). However, experimental ERB data indicates that animals have many more critical bands (Greenwood, 1961, 1990).

One other consideration when trying to determine the number of filters to use in the filter bank is that each filter should span at least $2\Delta f$, where Δf is the resolution of the spectral estimate, to compute an accurate value for the filter energy. To satisfy this constraint, the following inequality must hold:

$$n_f < \frac{2\left(F_p\left(f_{high}\right) - F_p\left(f_{low}\right)\right)}{F_p\left(f_{low} + \gamma / w_s\right) - F_p\left(f_{low}\right)} - 1, \tag{4.15}$$

where γ is the number of points desired in the lowest frequency filter + 1, w_s is the window size in seconds, f_{high} is the highest frequency included in the filter bank, and f_{low} is the lowest

frequency in the filter bank. The derivation of this equation is in Appendix C. Unfortunately, this maximum number of filters often causes the filters to be spaced more than one critical band apart for many species.

The shape of the masking filters has less effect on the classifiers used in this study, but shape is still an important consideration from a psycho-acoustic standpoint. Triangular filters were used in this study for computational simplicity, but more complex filter shapes such as those derived by Schroeder (1977) or Patterson *et al.* (1982) could be used as well. These more complex filter shapes are based on human acoustic data, and the applicability of these filter shapes to other species is largely unknown since there is little data on critical band masking filter shapes of non-human species.

Once the filter bank has been constructed, the filter energies are calculated using

$$\Theta[i] = \sum_{\omega} P[\omega]\Psi_i[\omega], \qquad 4.16$$

where $\Psi_i[\omega]$ is the ith filter's magnitude function and $\Theta[i]$ is the ith's filter's energy. The set of Θ energies represents a frequency warped, smoothed, and down-sampled power spectrum.

Equal Loudness Normalization

The next few components of the gPLP feature extraction model compensate for various psychoacoustic phenomena. Equal loudness normalization compensates for the different perceptual thresholds at each frequency for a species reflected in the audiogram. Hermansky (1990) originally used a function based on human sensitivity at the 40-dB absolute level derived using filter design theory. Since specific sensitivity curves are not available for many species, we present an alternative approach based on audiogram data.

A T-dB threshold curve can be approximated from the audiogram using

$$E[f] = -(A[f] - T), \tag{4.17}$$

where $A[f]$ is the audiogram data in decibels. It is generally accepted that 60dB is the hearing threshold for terrestrial species, while 120dB is the threshold for aquatic species (Ketten, 1998). This difference is the result of different reference pressures in water and air as well as the propagation differences between the two mediums (Ketten, 1998). The function $E[f]$ can be approximated by an nth order polynomial fit, $\hat{E}(f)$, for the purpose of interpolation. To better fit the polynomial to the data, it is strongly suggested that $E[\log(f)]$ be fit by a polynomial instead of $E[f]$ since audiogram data is often measured in equal log frequency steps. A 4th order polynomial is usually sufficient to accurately model the curve if log frequency is used because of the typical shape of the audiogram plotted on a logarithm of frequency axis. The constraint that $\hat{E}(f)$ not be negative is maintained by setting all negative values to zero. The equal loudness curve is applied by multiplying it by the filter bank energies:

$$\Xi[i] = \Theta[i]\hat{E}(cf_i). \tag{4.18}$$

Intensity-Loudness Power Law

The next component of the gPLP feature extraction model is to apply the intensity-loudness power law to $\Xi[i]$, the set of equal loudness normalized filter bank energies. Stevens (1957) formulated the law when he found that the perceived loudness of sound in humans is proportional to the cube root of its intensity. Although this exact relationship may not hold in other species, it is probable that because of the auditory system's structural similarity, a similar relationship exists between intensity and perceived loudness. Therefore, the following operation is performed on the normalized filter bank energies:

$$\Phi[i] = \Xi[i]^{0.33}. \tag{4.19}$$

Regardless of whether this relationship is exact, it compresses the dynamic range of the filter bank energies making it easier to model them by a low-order all-pole autoregressive model in the next analysis step.

Autoregressive Modeling

The rest of the components of the gPLP feature extraction model are associated with making the calculated features mathematically efficient and robust. The main purpose of autoregressive modeling is to reduce the dimensionality of the filter bank energies and smooth the spectral envelope. In this step of the gPLP feature extraction model, the filter bank energies, $\Phi[i]$, are approximated by an all-pole filter model of the form

$$H(z) = \frac{1}{\left(1 - p_1 z^{-1}\right)\left(1 - p_2 z^{-1}\right)\cdots\left(1 - p_n z^{-1}\right)} \qquad 4.20$$

where n is the order of the filter. The filter is derived using the autocorrelation method and the Yule-Walker equations as derived by Makhoul (1975). For more information on filter design using the autocorrelation method and linear prediction, which is mathematically equivalent, see Haykin (2002:136). The spectrum of the derived filter maintains the spectral peaks and valleys as represented in $\Phi[i]$, but represents the spectrum with many fewer coefficients. Autoregressive coefficients can also be converted to cepstral coefficients, which provide a number of computational benefits over filter bank energy representations.

The order of the LP analysis needed to capture the relevant peaks and valleys of the spectrum varies depending on the application. In general, $(n\text{-}1)/2$ peaks can be modeled by an nth-order all-pole filter. Hermansky (1990) found fifth-order filters to be appropriate because it was desirable to model the first two formants of human speech which can be used to uniquely define all English phonemes. By not using a higher order filter, the third and

fourth formants, more dependent on individual speaker variation, could be discarded from analysis. This was appropriate for the task of speech recognition.

However, animal vocalizations can have more harmonics and formants than human speech, therefore higher order filters are required to model this additional complexity. Lower order filters simply drop these upper harmonic peaks and model the strongest harmonics. Based on the vocalizations being analyzed and how they compare with the other classes of vocalizations, using a higher order filter can be a benefit or detriment depending on whether it is advantageous to model the upper harmonics. For example, a speaker identification task might benefit from modeling higher frequency formants and harmonics since they contain more vocal tract information. However, a call-type classification task might benefit from modeling fewer harmonics to ignore speaker-dependent spectral information.

Cepstral Domain Transform

The final component of the gPLP feature extraction model is to transform the autoregressive coefficients calculated in the previous step into cepstral coefficients. This transform is mathematically beneficial because Euclidean distance is more consistent in the cepstral domain than when used to compare autoregressive coefficients (Deller *et al.*, 1993). The autoregressive coefficients are transformed to cepstral coefficients using the recursion

$$c_n = -a_n - \frac{1}{n}\sum_{i=1}^{n-1}(n-i)a_i\, c_{n-i}\,, \tag{4.21}$$

where a_n are the autoregressive coefficients.

Liftering can be applied to the coefficients using

$$c'_n = \left(1 + \frac{L}{2}\sin\frac{\pi n}{L}\right)c_n\,, \tag{4.22}$$

where L is a parameter usually defined to be n or slightly greater than n (Deller *et al.*, 1993:378). Although this operation normalizes their magnitudes, it is non-linear and gives greater weight to the coefficients near index $L/2$. By giving greater weight to these coefficients, liftering gives more weight to the finer details of the spectrum. Liftering is also useful for visualization purposes.

Summary of gPLP Feature Extraction Model

The gPLP feature extraction model can incorporate experimental information from the species under study to generate perceptually relevant features from vocalizations. The features are also computationally efficient since a small number of coefficients can adequately represent the vocalization. To show the effects of the various analysis steps, and how the gPLP feature extraction model represents vocalizations, some examples are presented in the next subsection.

Applicability to MFCC Feature Extraction Model

Even though the MFCC feature extraction model does not use equal loudness curves, the filter bank adjustments discussed above can be used to create a generalized MFCC feature extraction model. The Greenwood warping function can replace the Mel-scale in the determination of the position of the filter banks. The number of filters to use in the filter bank can also be adjusted based on ERB data. Although MFCC analysis does not incorporate as much perceptual information as PLP analysis, it takes less computation time and therefore is sometimes more desirable. During implementation of the feature extraction models, the filter bank is typically calculated only once. Therefore, these changes can be incorporated with little increase in computation time if MFCC analysis is preferred.

Examples

The coefficients calculated by the gPLP feature extraction model can be visualized by displaying the spectrum of the all-pole autoregressive filter. As mentioned above, the cepstral conversion is primarily for computational purposes and therefore can be eliminated for these examples. The spectrograms of the filter can be thought of as a perceptual representation of the traditional spectrogram. A spectrogram represents the frequency content of a signal over time. Time is the horizontal axis while frequency is the vertical axis. At point in time where there is a large amount of energy at a particular frequency, that portion of the spectrogram is black while frequencies not contributing to the signal at that time are white. The spectrogram energies are scaled to make black represent the maximum frequency energy in the signal and white represent no frequency energy. Values in between are grey-scaled in this dissertation.

Figure 4.4 shows traditional FFT-based spectrograms of two African elephant vocalizations used in the experiments in chapter 5 along with the gPLP representation using 5^{th} order filters in the middle row and 18^{th} order filters in the bottom row. The equal loudness curve and filter bank are based on the perceptual abilities of the Indian elephant. The exact curve and filter bank are discussed in Chapter 5.

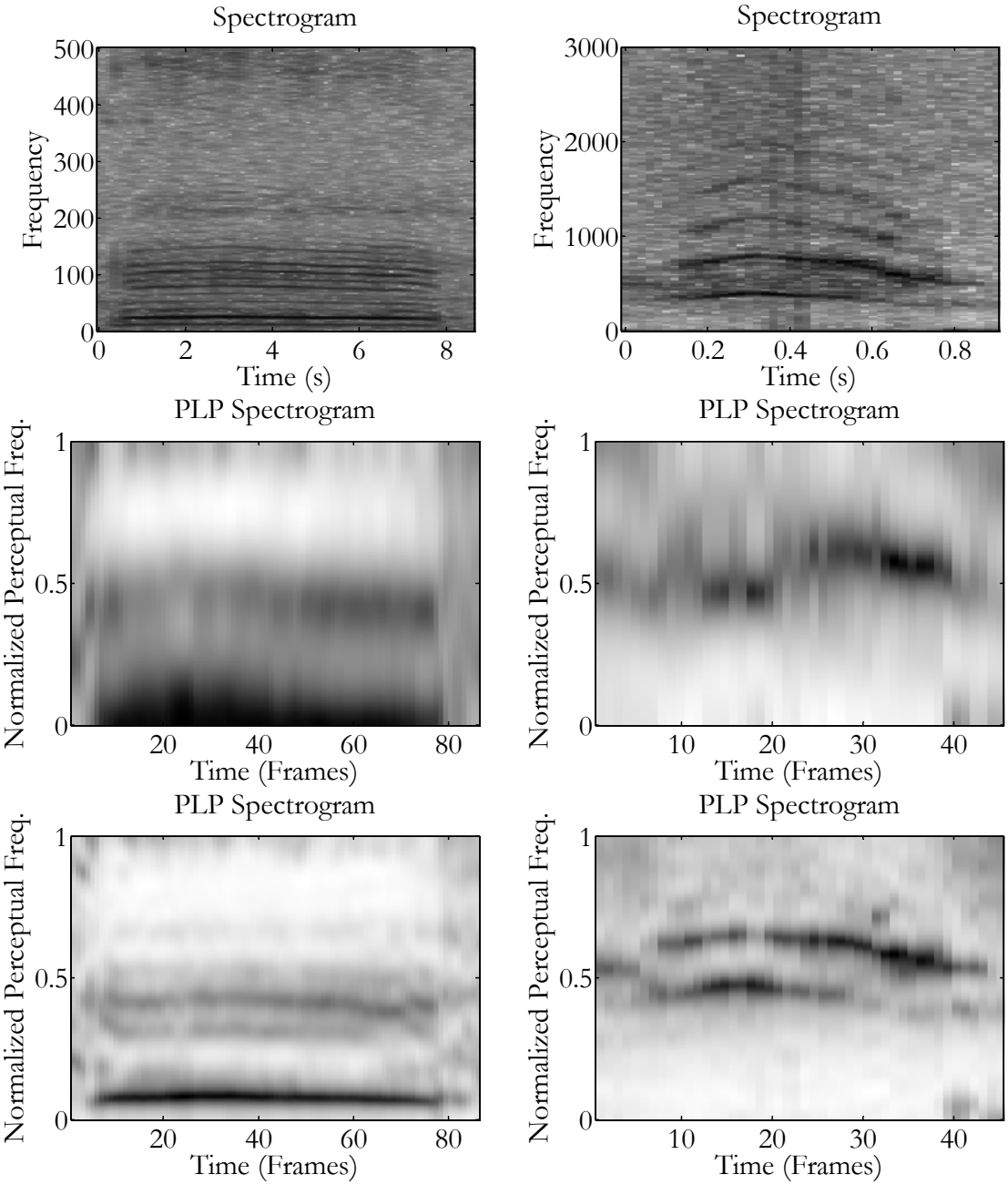


Figure 4.4 – gPLP Spectrograms of Elephant Vocalizations
Left: Elephant Rumble, Right: Elephant Trumpet
Dataset2\1R1_D3806RA and Dataset5\TR_2117MA

Although Hermansky (1990) showed that a fifth order filter was sufficient to model the human speech spectrum, these examples show that a filter of this size is inadequate. The African elephant vocalizations are clearly better represented by the 18th order filters which model the formant and harmonic structure more clearly. The lowest formant bleeds into the

0Hz range in the 5th order representation of the left vocalization and the formants and harmonics are not as distinctive in the 5th order filter representation. The formants are more clearly defined with darker formant regions and smaller formant bandwidth in the 18th order filter representations. The classification results that follow also show much higher classification accuracies when an 18th order filter is used.

The first vocalization, on the left, was included to show the ability of gPLP to model formant structure. This vocalization has two strong formants, one below 60Hz and the other near 100Hz. There is a third, weaker formant between 200Hz and 260Hz. These formants can be seen in the FFT-based spectrogram as the harmonics, the dark horizontal lines spaced about 12Hz apart, get darker at formant peaks and lighter at valleys in the spectral envelope. The 18th order perceptual spectrogram shows all three of these formant peaks at 0.1pHz, 0.4pHz and 0.7pHz. The perceptual spectrogram also smoothes out the harmonics shown in the FFT-based spectrogram.

The second vocalization, on the right, was included to show the ability of gPLP to capture quickly changing spectral characteristics and harmonics because of its frame-based structure. The 18th order filter was able to model the dynamics of both the strongest harmonic near 700Hz, and the harmonic near 500Hz shown by the downward curving black lines in the spectrogram and perceptual spectrograms. The bandwidth of these harmonics is also captured as can be seen at the end of the vocalization when the thickness of the strongest harmonic increases. Faster changing spectral characteristics can be modeled by reducing the frame step size.

The effect of the Greenwood warping can also be seen in the perceptual spectrograms of these vocalizations. The lower frequencies occupy a much larger range of the perceptual spectrograms. In the first vocalization, the second formant is at the lower quarter of the

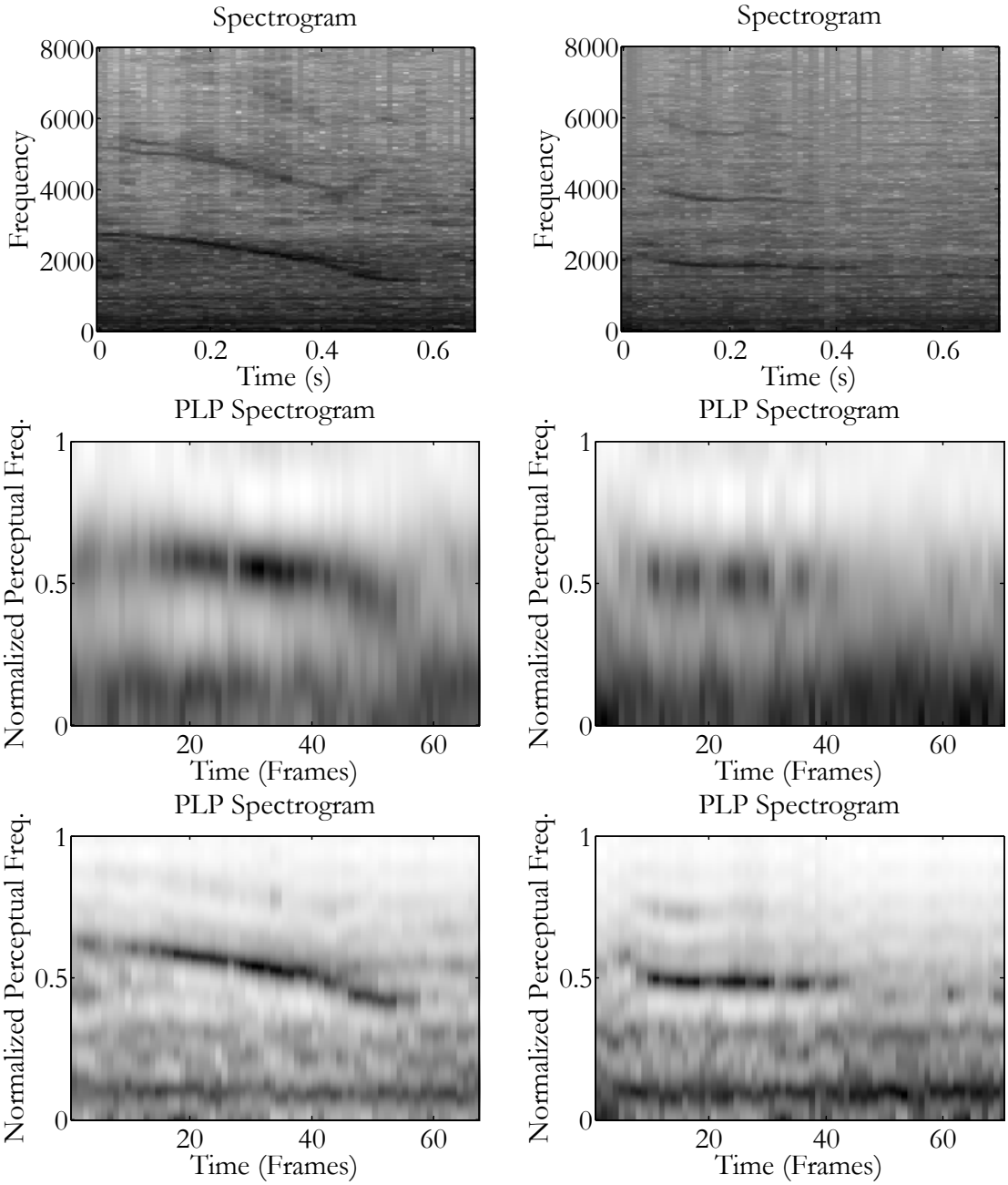


Figure 4.5 – gPLP Spectrogram of Beluga Whale Vocalizations
Left: Beluga Down Whistle, Right: Beluga WhineA
Set 1\dwnwhisa5 and Set 1\al6-01t1whinea

FFT-based spectrogram, but in the perceptual spectrogram, the second formant falls closer to the middle. This same effect can be seen in the second vocalization by comparing the location of the strongest harmonic.

The perceptual spectrograms also show the effect of incorporating the equal loudness curve. This is most apparent in the second vocalization where the lower portion of the perceptual spectrogram carries no significant energy. In addition to reducing the effects of noise outside the audible range of the species under study, it also focuses the analysis on the portion of the vocalization that can be heard the best. The effect is less obvious in the first vocalization because it focuses on a much smaller range of the elephant's hearing range.

Figure 4.5 shows the FFT-based spectrograms of two beluga whale vocalizations from the dataset used in chapter 6 along with the gPLP representation using 5^{th} and 18^{th} order filters. The equal loudness curve and filter bank used to calculate the gPLP coefficients are discussed in detail in chapter 7. As with the African elephant vocalizations, the 5^{th} order filter fails to capture the harmonic information present in the beluga whale vocalizations. However, the 18^{th} order filter captures the fundamental frequency as well as the first harmonic contour of both vocalizations.

In both of these vocalizations, the large amount of background noise is manifested in the FFT-based spectrograms by the rather dark background. However, the perceptual spectrograms have a much lighter background indicating the ability of the gPLP feature extraction model to filter out background noise and emphasize the spectral energy associated with the vocalization. The effect of the Greenwood warping is also evident from the plots, although to a lesser degree than in the African elephant vocalizations.

These examples show how the gPLP feature extraction model processes the waveform to generate a spectral view that incorporates information about the perceptual abilities of the species under study. Although gPLP coefficients are best suited for classification models that can model time-sampled data, they can also be used in statistical hypothesis tests. The

following section outlines the method used in this research to apply gPLP coefficients to statistical hypothesis testing.

Statistical Hypothesis Testing

Traditional bioacoustics signal research relies on statistical tests to support research hypotheses. Some of these hypotheses involve defining repertoires by showing that vocalizations are acoustically different or demonstrating that the individual making a vocalization can be determined by showing that vocalizations from two individuals are different. These traditional studies typically use features that are calculated over the entire vocalization from spectrograms such as maximum frequency, average fundamental frequency and duration. However, gPLP coefficients can also be used to conduct these statistical tests even though they are frame-based features.

One way to use gPLP coefficients in statistical tests is to treat the entire or most of the vocalization as a single frame and calculate the gPLP coefficients over this large frame. This approach has been done with cepstral coefficients (Soltis *et al.*, 2005). Treating the entire vocalization as a single frame, however, is not recommended because animal vocalizations, like speech, represent the output of a time-varying system. Therefore, the spectral characteristics of vocalizations are constantly changing throughout the duration of the vocalization. Calculating features over the entire vocalizations averages out the dynamics and this information about the dynamics is lost. The gPLP coefficients can also calculated using the centermost frame of the vocalization. Although this solves the issue of stationarity over the analysis window, this technique still fails to capture the dynamics of the vocalization.

Instead, the gPLP coefficients should be calculated as frame-based features as outlined in chapter 4. The difficulty with this is that each vocalization generates a number of data

vectors for use in the statistical test. Each data vector represents the vocalization at a different time within the vocalization. Because the vocalization is time-varying, the data vectors cannot be considered for a repeated measures statistical test. To overcome this problem, the data vectors need to be grouped into independent groups based on where they occur in the vocalization.

To perform this grouping, an HMM is trained for each class using all of the vocalizations. The feature vectors from each frame of the vocalization are aligned according these models using the Viterbi algorithm (Forney, 1973), and the state that each frame is aligned to becomes a second independent variable. Therefore, each statistical test has two independent variables, the class label and the state label which represents where in the vocalization the data vector occurred.

Analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA) are commonly used statistical tests to determine whether multiple classes of data originate from significantly separate distributions. MANOVA can also provide information about each dependent variable and the amount its distribution differs between classes. Both methods are commonly used in bioacoustics to show the differences between vocalization types.

Summary

This chapter has outlined the components of the gPLP framework, namely the gPLP feature extraction model, and the various classification models that can be used. The specific implementation details of the models were presented as well as the places where species specific perceptual information can be incorporated. These classification and feature extraction models together provide a framework for the analysis of animal vocalizations. The next two chapters give classification results using the gPLP framework presented in this chapter for two different species over various classification tasks.

Chapter 5

SUPERVISED CLASSIFICATION OF AFRICAN ELEPHANT VOCALIZATIONS

One application of the generalized perceptual linear prediction (gPLP) framework is as a supervised classification system. A supervised classification task is one in which a set of data has been labeled with the correct classification. The labeled data, called the training set, is then used to train the system to classify unknown data items, the test set. The dataset used in these experiments is a set of African elephant (*Loxodonta africana*) vocalizations. Elephant vocalizations were chosen for this experiment because researchers have studied the species for a number of years, especially their conspecific acoustic communication (Berg, 1983). The data is labeled with behavior annotations, the type of vocalization, the individual making the vocalization, and for females, estrous cycle information.

A number of classification tasks are investigated in this study. The first task is to determine the type of vocalization given a repertoire. The second task is to identify the elephant speaking. The third is to determine the estrous cycle phase of a female based on her rumble. The fourth task explores whether rumbles given in different contexts can be discriminated. While the first task uses vocalizations of all types, the last three tasks focus exclusively on rumbles. We discuss each of these tasks in turn.

Call Type Classification

Berg (1983), Poole *et al.* (1988), and Leong *et al.* (2002) have all used various schemes to categorize the repertoire of the African elephant which includes vocalizations with infrasonic content. Although there are slight differences between each categorization scheme, all agree that there are approximately 10 different sound types. FFT-based spectrograms of five of

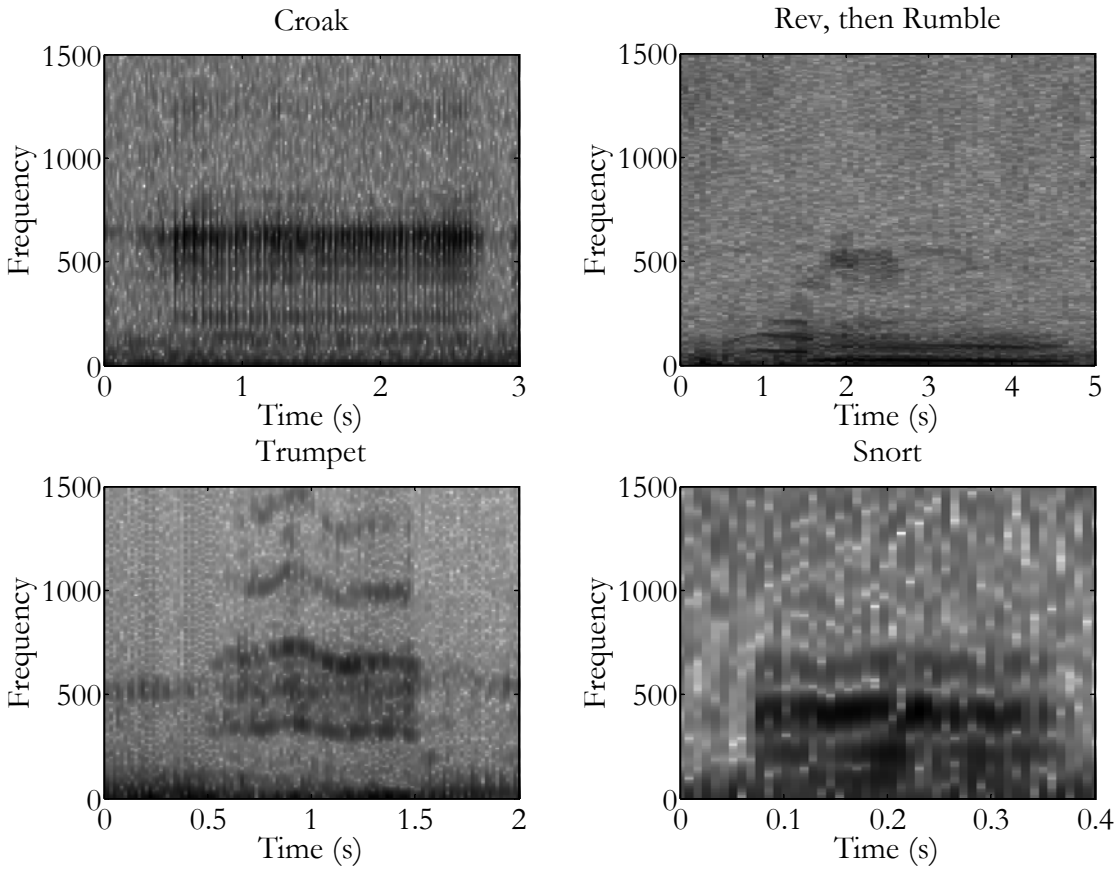


Figure 5.1 – African Elephant Vocalizations
Top Left: Croak, Top Right: Rev, then Rumble
Bottom Left: Trumpet, Bottom Right: Snort

the most common types of vocalizations are shown in Figure 5.1. The rumble, the most common vocalization, lasts for approximately 3 - 4 seconds and can have a fundamental frequency as low as 12 Hz. This vocalization is used for most conspecific communication. The low frequency characteristics of the rumble allow it to be heard over long distances. This has been verified through playback experiments (Langbauer Jr. *et al.*, 1991; Langbauer Jr. *et al.*, 1989; Poole *et al.*, 1988).

Other vocalizations include the rev which is usually followed by a rumble. The rev is made when the animal is startled. The croak usually comes in a series of two or three vocalizations and is commonly associated with the elephant sucking water or air into the trunk. The snort is a short, higher frequency vocalization that is used for a low-excitement

Name	Gender	Estimated Date of Birth	Call Type	Identity	Behavioral	Estrous
Robin	F	1970	6	34	9	27
Bala	F	6/1979	26	20	10	40
Fiki	F	6/1979	11	30	2	0
Thandi	F	3/1981	4	28	13	0
Moyo	F	10/1981	6	17	5	31
MacLean (Mackie)	M	1982	21	14	0	0

Table 5.1 – African Elephant Subjects and Number of Vocs. Used In Each Task

greeting. The familiar trumpet is most commonly produced when the elephant is very excited. The cry, growl, roar, and bark are other less frequently used vocalizations whose purpose is unclear.

This task is performed using both the MFCC model and the gPLP model to extract features. The HMM was primarily used for classification in all experiments to compare the performance of gPLP derived features to the performance of MFCC derived features. The DTW model is used to show the performance benefits of the HMM model.

SUBJECTS

One adult male and five adult female African elephants are the subjects of this experiment. They are housed at Disney's Animal Kingdom,™ in Lake Buena Vista, FL, and are part of a population of 13 elephants which exhibit many of the social dynamics seen in wild African elephants such as family bond groups. They are part of a long-term study of elephant communication which incorporates behavior, acoustic, and hormonal data to provide information about reproductive strategies. The birth of two elephants on 5/24/2003 and 7/6/2004 are an indication of some of the success of the program. The number of vocalizations in each dataset from the subjects is given in Table 5.1

DATA COLLECTION

The elephants are fitted with a custom-designed radio collar designed by Walt Disney World Co. Instrumentation Support Division of Ride and Show Engineering. Each collar

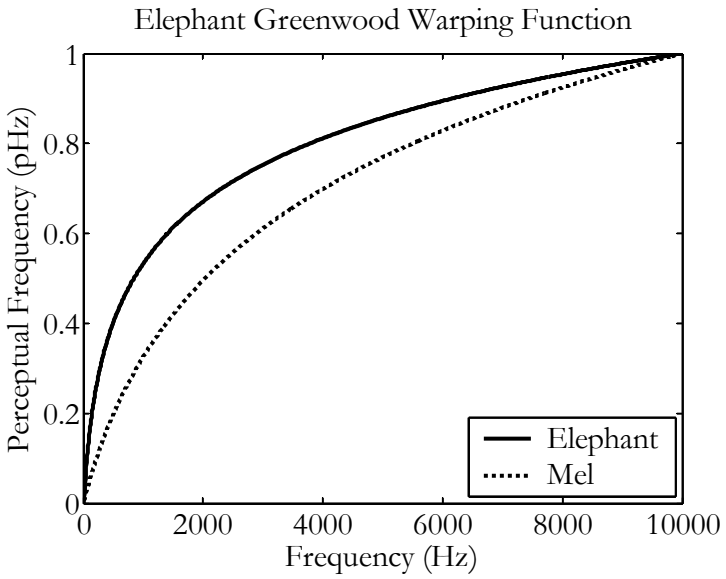


Figure 5.2 – Elephant Greenwood Warping Curve

contains a microphone and RF transmitter which transmits to the main elephant barn where the acoustic data from each collar is recorded on a separate channel of a TASCAM DA-38 8-channel DAT recorder made by TAEC America Inc. of Montebello, CA.

All of the vocalizations were manually extracted from the DAT tapes using Real-Time Spectrogram 2.0 (RTS) software made by Engineering Design of Belmont, MA. The vocalizations were passed through an anti-aliasing filter and stored on a computer, uncompressed, at a sampling rate of 7519Hz. Each vocalization visually identified on the spectrogram or acoustically identified was stored as a separate file. The nature of the data collection procedure led to a number of potential noise sources including RF interference, noise from passing vehicles, and the elephants submerging the collar in water or mud. All vocalizations were also amplitude scaled to normalize their power.

Feature Extraction

Features were extracted using a 60ms Hamming window, about three times larger than the window size typically used in human speech. The window size was lengthened to account for the lower fundamental frequencies of the African elephant vocalizations in

Frequency (Hz)	16	20	31.5	63	125	250	500
Hearing Threshold (dB re 20μP)	65.0	52.5	43.5	38.5	36.5	24.5	24.0
Frequency (Hz)	1000	2000	4000	8000	10000	12000	14000
Hearing Threshold (dB re 20μP)	8.0	19.0	23.0	42.0	55.5	72.0	79.0

Table 5.2 – Indian Elephant Audiogram Data

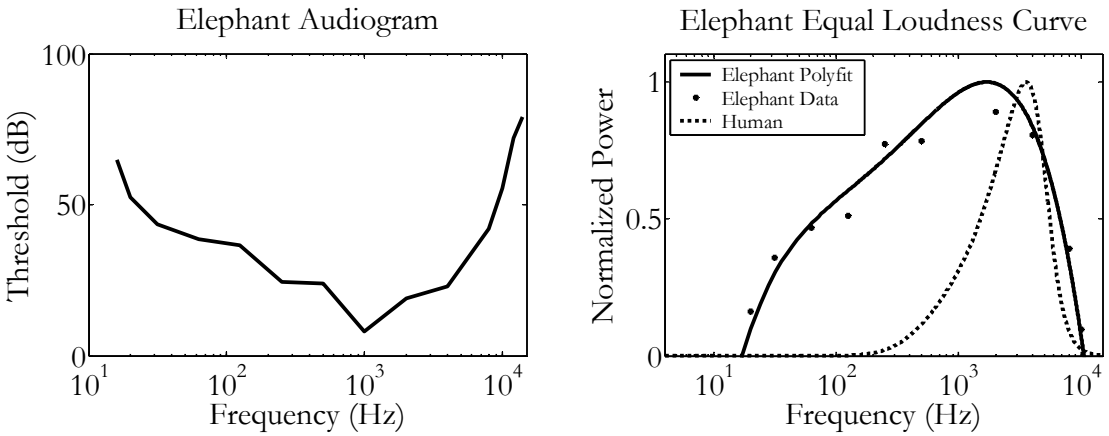


Figure 5.3 – Indian Elephant Audiogram and Equal Loudness Curve

comparison to human speech. The windows step size was 20ms, one-third the frame size giving a frame overlap of two-thirds.

The appropriate Greenwood frequency warping constants were calculated using 10Hz – 10kHz as the approximate hearing range of the elephant (Békésy, 1960; Heffner and Heffner, 1982). The appropriate equations from chapter 4 were used as follows:

$$k = 0.88, \tag{5.1}$$

$$A = \frac{F_{\min}}{1-k} = \frac{10}{1-0.88} = 83.333, \text{ and} \tag{5.2}$$

$$a = \log_{10}\left(\frac{F_{\max}}{A} + k\right) = \log_{10}\left(\frac{10000}{83.333} + 0.88\right) = 2.082. \tag{5.3}$$

Frequency (Hz)	250	500	1000
Actual ERB (Hz)	2.5	5	16
Call Type 30 filters, 10Hz -1500Hz	66	116	204
Speaker Identification 50 filters, 10Hz -500Hz	25	44	NA
Estrous and Behaviour 40 filters, 10Hz -300Hz	12	NA	NA

Table 5.3 – Approximate Filter Widths for Elephant Experiments

The resulting warping curve, using the above constants, was compared with the Mel scale shown in Figure 5.2. In this figure, the Mel scale has been normalized to the same range as the elephant Greenwood warping curve for comparison. Since 10Hz is used as $F_{\min}$, the lower bound of the filter bank must be greater than 10Hz, otherwise negative values of f_p will result. The reason for this constraint was explained in chapter 4.

The equal loudness curve was derived using the audiogram method outlined in chapter 4 using data from Heffner and Heffner (1982). The fourth-order polynomial fit of the audiogram data in Table 5.2 using the function polyfit in MATLAB® is

$$\hat{E}(f) = \begin{array}{c} -3.5056\times10^{-1} f^4 + 7.6473\times10^{0} f^3 - 6.1806\times10^{1} f^2 + \\ 2.2725\times10^{2} f - 2.9891\times10^{2} \end{array}. \qquad 5.4$$

The equal loudness curve and audiogram are shown in Figure 5.3.

The call-type classification task is performed a number of times using various feature extraction parameters to show the effect of each parameter on classification accuracy. Eighteen coefficients was experimentally determined over a large number of trials to give consistently better classification accuracies than different numbers of coefficients. Therefore, eighteen coefficients will be used whenever possible. The filter bank ranges used were also experimentally determined over a large number of trials.

Every effort was made to have the filter widths in each experiment be near 1 elephant ERB. However, as discussed in chapter 4, this is usually not possible because of the extremely small ERBs of most land mammals. Table 5.3 shows the experimentally determined ERB at various frequencies for an Indian elephant and the approximate filter widths at the same frequencies for the various filter banks used in these experiments. From the table, it can be seen that the filter bandwidths do not approach the actual width of the elephant ERB, but due to the limitations of Fourier spectral estimation it is impossible to make the filter widths smaller.

The specific experiments shown in the results sections are presented to show trends and demonstrate the effectiveness of various modifications to the models. They do not represent all of the experimental results acquired in this research. The details of the other parameters will be discussed in the results section for each experiment.

Model Parameters

The DTW system was implemented in MATLAB® using no global path constraints and the local paths as discussed in chapter 4. For the HMM classification systems, a three state left-to-right HMM was used to model each call type. This topology was chosen because each vocalization was relatively stationary throughout with the exception of trumpets, which tend to have extensive frequency modulation. HMMs with more states were tested, but this modification did not alter the classification accuracy to a significant extent. A three state silence model was added to the beginning and end of all vocalizations. All state observations were modeled with a single multivariate Gaussian. The Hidden Markov Model Toolkit 3.2.1 (HTK) from Cambridge University (2002) was used, with modification, to implement the feature extraction and HMM functionality.

		Speaker					
		Bala	Fiki	Mackie	Moyo	Robin	Thandi
Call Type	Croak	1	3	13	0	0	0
	Rumble	10	1	0	0	0	0
	Rev	11	0	2	0	0	1
	Snort	3	4	6	0	2	2
	Trumpet	1	3	0	6	4	1

		Speaker					
		Bala	Fiki	Mackie	Moyo	Robin	Thandi
Call Type	Croak	1	2	2	0	0	0
	Rumble	6	1	0	0	0	0
	Rev	7	0	0	0	0	0
	Snort	2	0	0	0	1	0
	Trumpet	1	3	0	5	4	0

Figure 5.4 – Call Type Distribution Across Speakers
Left: Complete Dataset; Right: Clean Dataset

RESULTS

The vocalization type classification experiment is analogous to an isolated word recognition task in human speech processing. Five different types of vocalizations, croak, noisy rumble, rev, snort, and trumpet, are classified in this experiment using a total of 74 different calls from 6 different elephants. The distribution of the 5 different calls across the speakers is shown in Figure 5.4.

Because of the rather small number of vocalizations, leave-one-out classification is used. In this experimental setup, all but one of the vocalizations is used to train the models which are then used to evaluate the one vocalization left out of the training set. Each vocalization is removed from the training set once, resulting in the models being trained once for each vocalization to classify each vocalization using models trained on the remainder of the data.

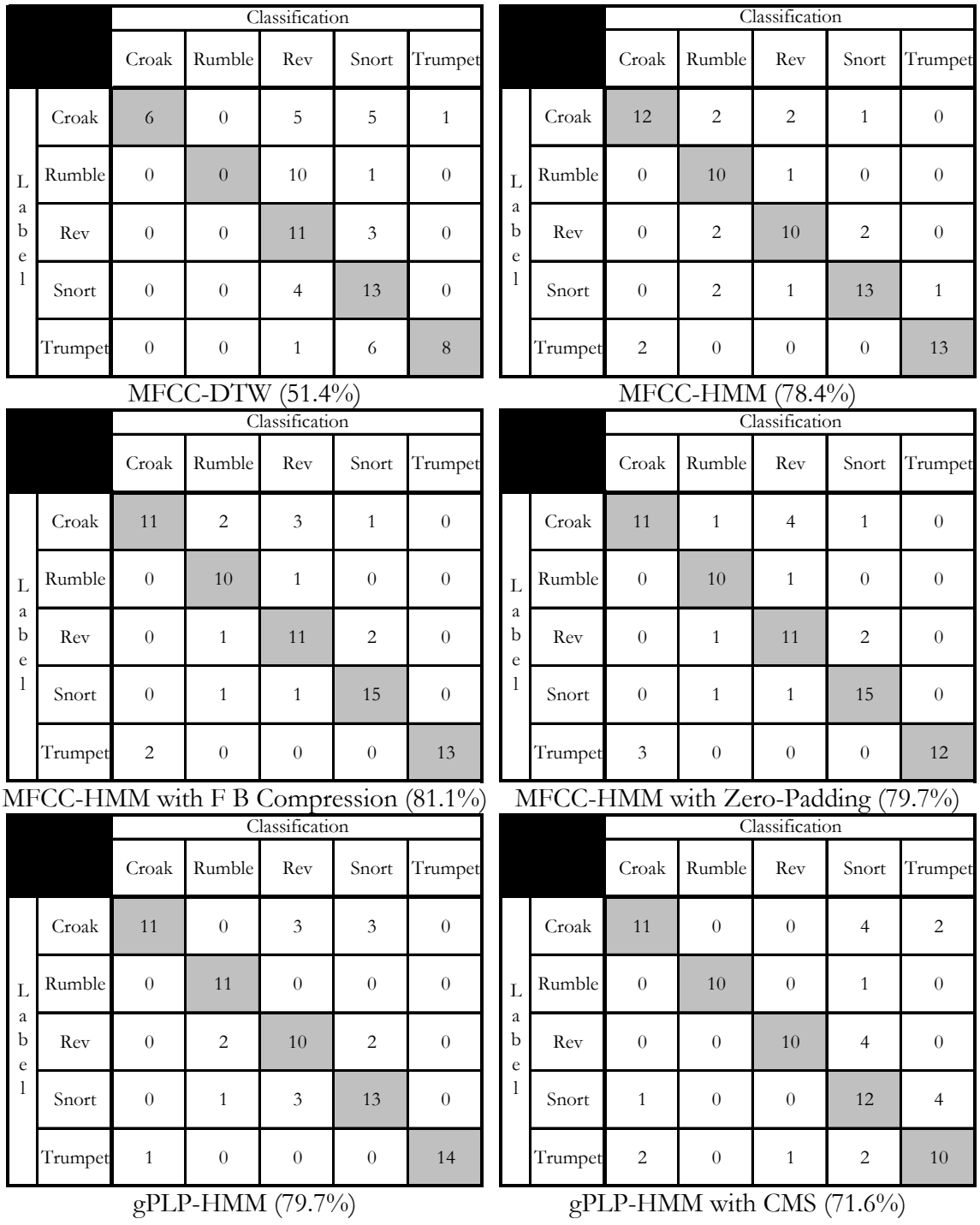


Figure 5.5 – Call Type Classification Results

The first classification system used for this task is a DTW classifier using eighteen MFCCs, derived using 30 filters spaced from 10Hz to 3000Hz. The pre-emphasis coefficient is set to 0.95 and log energy was included as a feature. These feature parameters

were chosen because they are similar to the parameters used in a human speech recognition system. The confusion matrix for this system is shown in Figure 5.5 at the top left. Each row represents the labeled class of the vocalization, while the columns represent the classification given to each vocalization by the system. Therefore, the numbers on the highlighted diagonal are the number of correctly classified vocalizations for each vocalization type.

The MFCC-DTW system was able to classify the vocalizations with an accuracy of 51.4%. Although this accuracy does not approach accuracies possible with human speech, it is well above the chance accuracy of 20%. As will be seen in the other classifications tasks, the DTW classification systems did not have good classification accuracies when compared to the HMM classification systems used later. However, the DTW systems did classify enough above the chance classification accuracy for further investigation to be warranted.

The classification matrix for this experiment shows that the DTW system classified a majority of the vocalizations as either revs or snorts, the shortest vocalizations. This is a good example of how the DTW classifier used in this study tended to favor one or two classes over the rest. The exact reason for this bias is uncertain, but it the bias was consistently present in many of the experimental trials using the DTW classifier. One hypothesis for this bias is that the DTW model favors shorter reference models which represent shorter vocalizations. Shorter reference models lead to a smaller DTW grid like the one shown in Figure 2.11. A smaller grid means that there are fewer local distances to accumulate along the recognition path and thus, smaller total path costs. Therefore, these shorter total paths across the grid cause the DTW model to favor shorter vocalizations.

The next experiment used a HMM classification system to compare the DTW classification system to an HMM system. The same features that were used in the last

experiment are used here to isolate the change in classification system. The classification accuracy of the MFCC-HMM system was 78.4% and the confusion matrix is in Figure 5.5 at the top right. Croaks, revs, and snorts are the most difficult call types for this system to classify. These are also the shortest vocalizations, meaning that they have fewer frames on which to base a classification decision. Rumbles were classified the best. This could be due to the fact that the rumbles are the longest vocalization or that they are the most distinct vocalization in comparison to the other types. It should also be noted that most of the rumbles came from a single speaker, while the snorts were distributed more evenly between speakers. Because the vocalizations were modeled single Gaussian state distributions, they are less likely to be able to model inter-speaker differences present in the snort data. This could account for the classification accuracy differences between rumbles and snorts.

To explore the effectiveness of compressing the filter bank range to the range of the vocalizations, the same experiment was run with the filter bank constrained to the range 10Hz to 1500Hz. The confusion matrix for this experiment is in Figure 5.5 at the top right. This system achieved an accuracy of 81.1%. Although the accuracy did increase, the improvement was minimal because the compression of the filter bank, in this case, does not affect the spacing of the Mel-frequency filters to a large degree. In the original spacing, there were only a few filters in the range 1500Hz to 3000Hz. Therefore, the compression only moved a few filters into the range of the vocalizations.

The next parameter that was adjusted was the amount of padding of the signal before the Fourier transform was calculated. Zero-padding the signal increases the frequency resolution through interpolation. In this experiment, the frame is zero-padded to four times its original length. The filter bank range is kept at 10Hz – 1500Hz. The confusion matrix for this experiment is in Figure 5.5 at the middle left. This system had a classification

accuracy of 79.7%. Zero-padding did not improve the classification rate in this case because the filter bank was not compressed enough in the low frequencies to cause a problem with calculating the filter energies as discussing in chapter 4.

This task was then performed using the gPLP feature extraction model. The filter bank range was set to 10Hz – 1500Hz and the signal was not zero-padded since in the previous experiment it was not beneficial. Eighteen coefficients were derived using 30 filter bank energies, just as in the MFCC experiments. The confusion matrix for this experiment is shown in Figure 5.5 at the middle right. The gPLP-HMM system achieved an accuracy of 79.7%, one classification worse than the MFCC-HMM system using the compressed filter bank range. The main difference was that the MFCC-HMM system did a better job of classifying snorts, a higher frequency vocalization. The greater weight the MFCC frequency warp gives to higher frequencies than the Greenwood warp of the gPLP model is probably the reason the MFCC-HMM system classified snorts better. It is interesting to note, however, that the gPLP-HMM system did classify the trumpets and rumbles better, which have been shown to be the more salient vocalizations. This experiment was also run with the pre-emphasis filter coefficient set to 0.0, making the pre-emphasis filter an all-pass filter. This parameter change did not affect the classification accuracy which stayed at 79.7%.

Finally, the task was performed using cepstral mean subtraction (CMS). CMS is a method used to normalize the effects of convolutional noise in a system. As discussed in chapter 2, a convolution in the time domain becomes a multiplication in the spectral domain and a subtraction in the cepstral domain. Therefore, by subtracting the mean cepstrum from the cepstral coefficients, convolutional noise can be removed assuming the noise is stationary throughout a vocalization. Cepstral mean subtraction is a commonly used technique in human speech processing. The confusion matrix for the gPLP-HMM system

with cepstral mean subtraction is in Figure 5.5 at the bottom right. This system had a classification accuracy of 71.6%, the worst of those presented. This system was especially poor at classifying snorts and trumpets. This poor performance could indicate that the noise present in the data set is additive as opposed to convolutional or that the noise is nonstationary.

Although the classification accuracies were significantly above chance, it was expected that they would be higher since these vocalization types can be identified relatively easily by experts. The wide range of frequencies used by the vocalizations and the short duration of the rev and snort are the main difficulties in performing the classification. The rumbles have very little energy above 300Hz, but most of the other vocalization's spectral energy is completely above 300Hz. This disparity creates difficulties in designing an effective filter bank.

It is also important to note that in most of the experiments presented; the difference in classification accuracy is only 2.7%, or 2 vocalizations (the cepstral mean subtraction experiment is not included). Because of the relative simplicity of this task, this could indicate that 80% accuracy is near a threshold of what can be reasonably expected given the quality of the data.

A major source of error in the classification is that many of the original 74 vocalizations are noisy or digitized over a small magnitude range resulting in large quantization error. Therefore, the same experiment was performed using only clean vocalizations. All vocalizations that had a scale factor of greater than 1100 during power normalization were considered poorly quantized and removed. To determine which vocalizations were noisy, a metric related to signal to noise ratio was developed. The energy in each frame of the

		Classification				
		Croak	Rumble	Rev	Snort	Trumpet
Label	Croak	4	0	1	0	0
	Rumble	0	7	0	0	0
	Rev	0	1	6	0	0
	Snort	0	0	0	3	0
	Trumpet	0	0	1	1	11

MFCC-HMM with FB Compression (88.6%)

		Classification				
		Croak	Rumble	Rev	Snort	Trumpet
Label	Croak	5	0	0	0	0
	Rumble	0	7	0	0	0
	Rev	0	1	6	0	0
	Snort	0	0	0	3	0
	Trumpet	2	0	0	0	11

gPLP-HMM (91.4%)

Figure 5.6 – Call Type Classification on Clean Data

vocalization was calculated and the signal to noise characteristic (SNC) for each vocalization was computed using

$$SNC = {FrameEnergy_{max}}/{FrameEnergy_{ave}}. \quad 5.5$$

Vocalizations with a SNC of less than 5.0 were removed from the dataset.

After removing noisy and poorly quantized vocalizations, 35 vocalizations were left in the dataset. The distribution of the vocalizations across the speakers in this reduced dataset is shown in Figure 5.4. This experiment was run using both the MFCC-HMM system and the gPLP-HMM system with the filter bank range compressed to 10Hz – 1500Hz. As in the previous experiments, eighteen coefficients were derived from 30 filter energies and log energy was included as a feature. Cepstral mean subtraction and zero-padded were not used. The classification matrix for this experiment is in Figure 5.6. The classification accuracies for this experiment are 88.6% and 91.4% for the MFCC-HMM and gPLP-HMM systems respectively. These are significantly higher than the accuracy using the entire data set. Revs still account for a number of the misclassifications along with trumpets. The addition of a more complex duration model could resolve this confusion since revs are typically much

shorter than the other vocalizations. A more sophisticated grammar could also resolve this problem since revs are usually followed by rumbles, which are easily classified.

These results show that the different types of vocalizations can be easily classified by the gPLP framework using a supervised HMM classification model. Although the accuracies are not as high as anticipated, noise and lack of a language model contributed to misclassifications. The next task will attempt to determine which elephant subject made each vocalization. Unlike the call type task, only rumbles will be used in this task, making it a call-dependent speaker identification task.

Speaker Identification

This experiment is designed to show that the individual making a vocalization can be identified by acoustics alone. Physiological differences in the each elephant's vocal tract such as shape and length should affect the parameters of the vocal tract filter of each elephant. This difference in filter parameters is reflected in the extracted cepstral coefficients. Therefore, by training a separate HMM for each elephant's vocal tract characteristics, the individual making the vocalization can be determined. Another reason for the individual difference in the vocalizations could be an effect similar to human dialects or accents. Dialects have been found in a number of studies (Dayton, 1990; Santivo and Galimberti, 2000), and since the origins of the subjects of this study are varied, the elephants could have developed geographically localized or family group accents or dialects.

In support of the hypothesis that the individual making the vocalization can be determined, there is evidence that African elephants can discriminate between the vocalizations of different elephants. McComb *et al.* (2000) showed that African elephants could tell the difference between vocalizations from familiar elephants and vocalizations from unfamiliar elephants through playback experiments. In addition, it was predicted that

the elephants in the study would have to be familiar with the vocalizations of at least 100 adult elephants for the elephants to make the observed discriminations. In another study, Soltis *et al.* (2005) was able to perform speaker identification from 6 different speakers using vocalizations collected from the same elephants used in this work. A discriminant function analysis (DFA) classifier, using 16 features ranging from formant locations and amplitude to basic fundamental frequency measures, was able to classify the vocalizations by speaker with an accuary of 60.0%.

This experiment was performed using three different classification systems. The first system is DTW-based, using traditional MFCC features. The second system uses MFCC features with HMM models, and the third system uses gPLP features with HMM models. After these results, a simple maximum likelihood classifier using traditional bioacoustic features extracted from a spectrogram is discussed for the sake of comparison.

Subjects

The vocalizations from the same adult male and five adult females that participated in the call type experiment were used in this experiment. See Table 5.1 for the number of vocalizations used from each subject.

Data Collection

The data was collected in the same manner as in the call type experiment. All vocalizations were amplitude scaled using variance normalization to normalize their power. The dataset consists of 143 rumbles made in various behavioral contexts from six different elephants. The DTW system was used early in the research, and therefore uses only a portion of the data (47 rumbles from 5 speakers), before the rest was acquired.

Feature Extraction

Both the MFCC and gPLP features were extracted using a 300ms Hamming window with 100ms frame step size for the HMM systems. A larger window and step size were used in this experiment in comparison to the call type experiment because rumbles are of a lower frequency than the other vocalizations. Rumbles are also longer. Therefore, even with the larger step size, an adequate number of frames can be analyzed. The same Greenwood warping constants and equal loudness curve used in the call type task were used in this experiment. As in the call type task, this task is run a number of times using different values for the feature extraction parameters.

Model Parameters

The DTW system was implemented in MATLAB using no global path constraints and the local paths as discussed in chapter 4. For the HMM experiments, a three state left-to-right HMM was used to model each class with an additional three state HMM to model the silence both following and preceding the vocalization. All state observations were modeled with a single multivariate Gaussian. More details are discussed in the call type experiment section.

		Classification					
		Bala	Fiki	Mackie	Moyo	Robin	Thandi
Label	Bala	5	0	2	10	2	1
	Fiki	2	13	0	11	2	2
	Mackie	0	0	8	3	2	1
	Moyo	1	0	2	9	4	1
	Robin	2	3	6	12	8	3
	Thandi	2	3	3	10	4	6

MFCC-DTW (34.3%)

		Classification					
		Bala	Fiki	Mackie	Moyo	Robin	Thandi
Label	Bala	9	2	0	3	3	3
	Fiki	3	15	0	3	0	9
	Mackie	0	0	10	2	2	0
	Moyo	3	0	2	3	7	2
	Robin	5	1	6	6	12	4
	Thandi	2	3	5	1	1	16

MFCC-HMM (45.5%)

		Classification					
		Bala	Fiki	Mackie	Moyo	Robin	Thandi
Label	Bala	15	1	0	0	4	0
	Fiki	0	28	0	0	0	2
	Mackie	1	1	10	1	0	1
	Moyo	1	1	0	12	1	2
	Robin	4	0	1	6	23	0
	Thandi	0	4	3	3	1	17

MFCC-HMM with F B Compression (73.4%)

		Classification					
		Bala	Fiki	Mackie	Moyo	Robin	Thandi
Label	Bala	18	0	0	0	2	0
	Fiki	1	24	0	1	0	4
	Mackie	0	0	13	1	0	0
	Moyo	1	0	1	14	1	0
	Robin	4	0	0	0	29	1
	Thandi	0	2	0	3	0	23

gPLP-HMM (84.6%)

Figure 5.7 – Speaker Identification Results

RESULTS

As in the call type classification experiment, leave-one-out classification was used for all experimental trials. The first system tested was a dynamic time warping (DTW) classifier using features extracted with the MFCC model. The feature extraction process used parameter values similar to those in a human speaker identification system. Using HTK 3.2.1, eighteen MFCCs, using 50 filters placed between 10Hz and 3000Hz, were extracted from the vocalizations along with log energy. A pre-emphasis coefficient of 0.95 was used and no zero-padding was performed on the signal. The confusion matrix for this experiment is in Figure 5.7 at the top right. The classification accuracy for the MFCC-DTW

system is 34.3%. This experiment shows the weakness of DTW in call-independent speaker identification. These rumbles were made in multiple behavioral contexts and therefore, probably mean different things. Hence, the DTW system has difficulty generalizing to the speaker because it is template-based and the rumbles were not consistent for each speaker. The template can only model one type of rumble for each speaker, thus, it models the average rumble for each speaker which may not be similar to any of the specific types of rumbles in the dataset.

The next experiment uses an HMM classifier implemented in HTK 3.2.1 (Cambridge University Engineering Department, 2002). The same feature extraction parameters used in the previous experiment were used here to directly compare the two different classification models. The classification accuracy of the HMM-MFCC system is 45.5% and the confusion matrix is in Figure 5.7 at the upper right. The HMM classifier performs better than the DTW system, probably due to the ability of the HMM to model the vocalizations in a much more flexible manner. The additional MFCCs and filters in the filter bank could also contribute to the improved classification accuracy.

Since rumbles have their energy concentrated below 500Hz, it is expected that filter bank compression will make a bigger difference in the classification accuracy in this task than in the call type classification task. To test this hypothesis, the experiment is run with the same parameters as the last, with the exception that the range of the filter bank is set to 10Hz – 500Hz. The confusion matrix for this experiment is in Figure 5.7 at the middle left. This system had a classification accuracy of 73.4%, much improved over the other two experiments which used a larger filter bank range. This experiment was performed with and without zero-padding the signal prior to the Fourier transform, but, as in the call type task, it had no effect on the classification accuracy.

The next experiment uses the gPLP model for feature extraction instead of the MFCC model. For the sake of comparison, the same parameter set as in the MFCC experiment is used. Eighteen gPLP coefficients are derived from 50 filter bank energies with the filter bank spaced between 10Hz and 500Hz. The confusion matrix for this experiment is in Figure 5.7 at the middle right. This experiment's classification accuracy was 84.6%, much higher than the previous experiments on this task. Other experimental runs confirm that the main reason for this performance increase is the Greenwood warping of the spectrum which weights the lower frequencies, where the majority of the rumble's energy is located, much greater than the Mel-scale. The same experiment was also run zero-padding the signal prior to performing the Fourier transform, but the improvement was only one more correct classification which increased the accuracy to 85.3%. Cepstral mean normalization had the same effect as in the call-type classification task, decreasing the accuracy significantly.

From the various confusion matrices, it can be seen that Thandi's vocalizations were the most difficult to classify, while Mackie's were the easiest. The two pairs of elephants that were confused the most often were Fiki/Thandi and Bala/Robin. From this information, it can be inferred that the elephants in these pairs have rumbles with similar acoustic characteristics. However, the confusions could also be due the paired elephants making a number of vocalizations with the same meaning.

Maximum Likelihood Classification with Spectrogram Features

The gPLP framework is capable of performing a speaker identification task with classification accuracies near 85% with six different possible speakers. To compare this performance with traditional bioacoustic techniques, spectrogram-based features were hand-extracted from 60 of the elephant vocalizations in the dataset, 10 from each speaker. These features include start fundamental frequency, end fundamental frequency, maximum

fundamental frequency, minimum fundamental frequency, fundamental frequency range, duration, and percent to maximum frequency. The percent to maximum frequency was calculated by dividing the time in seconds from signal onset to the maximum frequency by the total duration of the vocalization. The fundamental frequency was measured using the second harmonic in all vocalizations, so the fundamental frequency features measured were actually twice the fundamental frequency. This was done because the second harmonic was the strongest harmonic in each vocalization and the changes in the frequency contours were more apparent. Therefore, the measurements could be made with better accuracy and precision. Because an HMM is not appropriate for features measured only once per vocalization, a multivariate Gaussian was used to perform maximum likelihood classification.

The classification was performed using leave-one-out verification over a number of different parameter configurations. First, the covariance matrix was calculated on either a global basis, using all vocalizations except the test, or a class basis, using only the vocalizations from the speaker being evaluated. Second, because using a full covariance matrix sometimes generated a non-symmetric matrix or a matrix that was not positive definite, the use of a diagonal covariance matrix was explored. Finally, principal components analysis (PCA) was performed on the original features. The classification was performed using both the original features and various dimensions of the PCA-transformed features. The results are shown in Table 5.4. The text "Error" indicates that the standard MATLAB® covariance function generated non-symmetric matrix or that the matrix was not positive definite.

	Original Features	First 2 PCA Dimensions	First 3 PCA Dimensions	First 4 PCA Dimensions	All PCA Dimensions
Class-Specific Full Covariance	Error	35.0	38.3	38.3	Error
Class-Specific Diagonal Covariance	41.7	33.0	43.3	41.7	41.7
Global Full Covariance	Error	35.0	40.0	40.0	38.3
Global Diagonal Covariance	43.3	35.0	40.0	40.0	40.0

Table 5.4 – Maximum Likelihood Classification Results

These results show that while the results are above chance (16.7%), the accuracies are much lower than the classification accuracies accomplished with the gPLP framework. The accuracies are also much lower than the 60.0% from Soltis *et al.* (2005) which uses an additional 10 features and a DFA classifier. It is interesting to note that the best accuracy was achieved using the original features. This indicates that the PCA transformation does little to improve classification. It is also interesting to see that the best PCA classification is when the first three dimensions are included. This indicates that the last four PCA dimensions do not contribute to the classification accuracy, but only confuse the system.

Results from the speaker identification experiments show the improvement in classification accuracies as the feature extraction process incorporates additional species-specific perceptual information. While the original system, based on DTW and the MFCC feature extraction model performed poorly, the system was able to reach accuracies of over 80% using the gPLP feature extraction framework with an HMM classifier. Speaker identification experiments performed using traditional spectrogram-based features show that the gPLP framework greatly outperforms traditional spectrogram features on this task. The next task tries to determine the hormonal state of female elephants. It is hypothesized that females vary their vocalizations in order to attract males prior to ovulation.

Estrous Cycle Determination

The next task is designed to test the hypothesis that female elephants change their vocalizations just prior to ovulation to attract males for reproduction. The elephant estrous cycle is unique among animals. It consists of three distinctive phases instead of two. An 8 – 11 week long luteal phase is followed by an anovulatory period of follicular growth marked by a surge in luteinizing hormone. About a week after the anovulatory hormonal peak, a second surge in luteinizing hormone results in ovulation. The purpose of the anovulatory phase is uncertain (Leong *et al.*, 2003).

One hypothesis for the purpose of the anovulatory phase is to signal the female to attract males. Because low-frequency rumbles travel a long distance, these rumbles could be used to signal males many kilometers away that the female is soon ready for reproduction. The male would then have time to travel the distance to the female and arrive in time for successful reproduction. Classification of rumbles by the estrous phase of the female would indirectly support this hypothesis by showing that the rumbles do vary by estrous cycle phase. Whether the male elephants perceive or use these differences would have to be shown by playback experiments.

Subjects

The rumbles of three adult cycling females were used in this experiment. See Table 5.1 for the number of vocalizations used from each subject.

Data Collection

The data was collected in the same manner as in the call type experiment. All vocalizations were amplitude scaled using variance normalization to normalize their power. Rumbles are labeled with the phase of the estrous cycle in which it was made using luteal, anovulatory follicular and ovulatory follicular as the class labels. Only vocalizations made ±6

days from the midpoint of the luteal phase and ±6 days from each luteinizing hormone peak were used in this experiment resulting in 98 vocalizations from three cycling females.

To determine hormonal data for the estrous experiment, blood samples were collected twice per week during the luteal phase and daily during the follicular period from the three cycling females. Serum progestogen and LH concentrations were measured using enzyme immunoassays previously validated for African elephants (Graham *et al.*, 2002; Graham *et al.*, 2001). Details of the assay methodology can be found in Leong *et al.* (2003).

Feature Extraction

Features were extracted using a 300ms Hamming window with 100ms frame step size because all rumbles are being classified as in the speaker identification task. The same Greenwood warping constants and equal loudness curve used in the call type classification task were used in this task. These various parameter changes in each experiment will be discussed in detail in the results section.

Model Parameters

A three state left-to-right HMM was used to model each class with an additional three state HMM to model the silence both following and preceding the vocalization. All state observations were modeled with a single multivariate Gaussian unless otherwise noted. These are the same model parameters used in the call type experiment. More details are discussed in the call type experiment section.

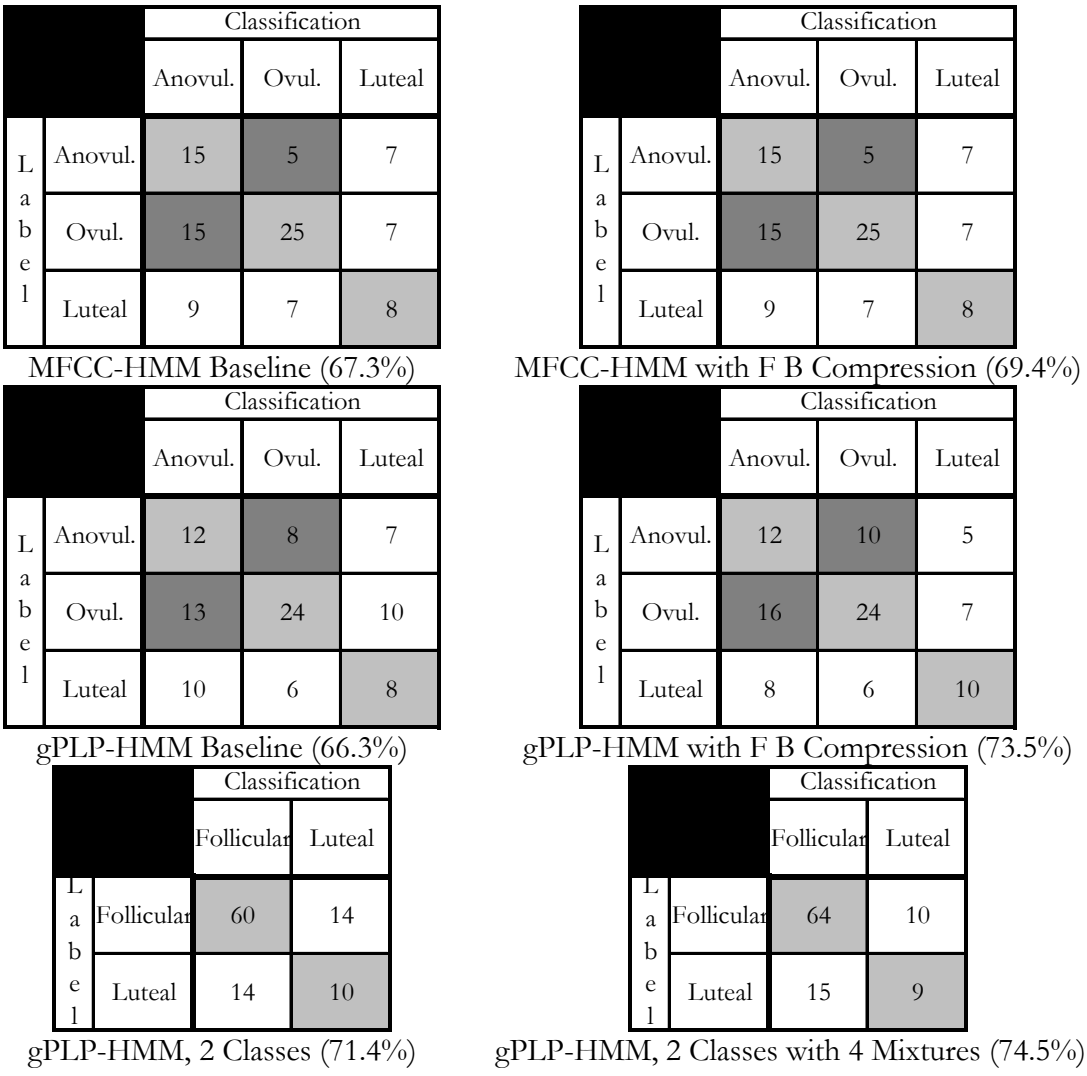


Figure 5.8 – Estrous Cycle Determination Results

RESULTS

The rumbles were classified using leave-one-out classification to maximize the use of available data in all experiments. The baseline MFCC-HMM system for this task uses eighteen coefficients extracted with 40 filters spaced between 10Hz and 3000Hz. The signal was not zero-padded, the pre-emphasis coefficient was set to 0.95 and log energy was included in the feature vector. The confusion matrix for this experiment is shown in Figure 5.8 at the upper left. Although the classification accuracy for this experiment is 50.0% if this

is considered a three-class problem, the accuracy is 67.3% if the two follicular classes, anovulatory and ovulatory are folded together. Folding together similar phonemes after classification is common in speech recognition experiments (Lee and Hon, 1989). It is appropriate here because the hypothesis is that the vocalizations change during the entire follicular phase of the estrous cycle to attract males before ovulation. As a result, the males can arrive at the female in time for copulation. The high number of misclassifications of ovulatory calls as anovulatory calls and vice versa supports this theory.

With the folding, not only are the diagonal cells in the classification matrix considered correct classifications, but the two darker shaded cells where the ovulatory and anovulatory phases intersect are also correct. It should be noted that folding the two follicular classes together also increases the chance classification to 55.6% since 5 of the 9 cells in the classification matrix are now correct classifications. This is in contrast to the 33.3% chance classification rate if the two follicular classes are not folded.

Since the rumble has most of its energy below 300Hz, the task is performed with the filter bank compressed to the range 10Hz – 300Hz. The frequency range was reduced further than in the speaker identification experiment for the purpose of noise reduction and to help the classifier focus on the low frequency characteristics of the vocalizations. The classification matrix of this experiment is in Figure 5.8 at the upper right. The classification accuracy is 49.0% with the follicular classes unfolded and 69.4% with the follicular classes folded. Unfortunately, compressing the filter bank did not lead to as large of an increase in classification accuracy as in the speaker identification task. One reason could be is that the feature set being used here does not incorporate some of the features of the rumble which differentiate the follicular rumbles from the luteal rumbles. Neither zero padding the frames nor cepstral mean subtraction affect the classification accuracy to a large degree.

The next experiment incorporates the gPLP feature extraction model. The baseline gPLP system uses 18 coefficients and 40 filters spaced between 10Hz and 3000Hz. These are the same parameters used for the MFCC-HMM baseline. The classification matrix for this experiment is in Figure 5.8 at the middle left. The classification accuracy is 44.9% with the follicular classes unfolded and 66.3% with the classes folded. This is only a classification difference of one vocalization when compared to the MFCC-HMM baseline. Therefore, it is hard to draw strong conclusions about which may be better.

To compare the gPLP feature extraction model to the MFCC model further, the filter bank range is compressed to 10Hz – 300Hz. As in the MFCC experiment, the signal was not zero padded and 40 filters were used along with 18 coefficients. The classification matrix for this experimental run is in Figure 5.8 at the middle right. The classification accuracy is 47.0% with the follicular classes unfolded and 73.5% with the classes folded. The higher classification accuracy when compared with a similar MFCC-HMM system is probably due to the closer spacing of the filters in the lower frequencies. Zero padding and cepstral mean subtraction, as in the MFCC-HMM systems, had little effect.

Since the two follicular classes were commonly misclassified, it was hypothesized that the rumbles made in these two phases of the estrous cycle are actually the same vocalization. To test this hypothesis, the two follicular classes were given a common label, and the task was performed as a two class classification problem. Eighteen gPLP coefficients and 40 filters spaced between 10Hz and 300Hz were used to test this hypothesis. The results of this experiment are shown in Figure 5.8 at the bottom left. The classification accuracy of this experiment was 71.4%. Given that the accuracy went down a small degree when the two follicular classes were given the same label, it is likely that these two vocalizations are

extremely similar, if not identical. However, the uneven number of examples for each class, 74 follicular compared to 24 luteal, makes the results less reliable.

Finally, to try to increase the accuracy more, the two class experiment was preformed again with four Gaussian mixtures in each HMM state. Refer to chapter 2 for a discussion of Gaussian mixture models. The mixtures were added with the intention of better modeling the slight differences of the rumbles between speakers and between the two follicular phases. The same gPLP feature extraction parameters as used in the previous two-class experiment are used here. The classification matrix for this experiment is shown in Figure 5.8 at the bottom right. The classification accuracy is 74.5%, the highest achieved for the task. This would indicate that the distribution of the coefficients for each type of vocalization is more complex than a single Gaussian. This complexity could be due to speaker differences, estrous phase differences, or even different recording conditions during the study.

One possible reason for the lower classification accuracies on this task than the two previous tasks is that the rumbles included in the dataset may not all be related to the female's estrous cycle. If the hypothesis that females alter their vocalizations during the follicular phases to attract males is true, it is reasonable to assume that not all of the vocalizations of the female during the follicular phase are meant for this purpose. Those vocalizations meant for other purposes would confuse the classification system since they really do not belong in the dataset. To build the classifier properly, the vocalizations would need to be labeled with the meaning of each vocalization, and with our limited knowledge of African elephant vocal communication, this is currently not possible.

The results of this task show that female elephants vary their vocalizations to a significant degree during follicular hormonal states. Whether males are able to perceive

these differences must yet be proven through playback experiments. The gPLP framework, using an HMM classification model was able to discriminate between vocalizations made during follicular and luteal hormonal states, but not between the anovulatory and ovulatory states. The next task is similar in nature, but tests the discrimination ability of the framework when applied to vocalizations made in different behavioral contexts. This task is the most comparable of the four tasks to human speech recognition, and thus, the most difficult.

Behavioral Context

The purpose of this experiment is to attempt to determine whether elephants use different rumbles to communicate different things. Since the acoustic data was recorded along with time-synchronized video, the acoustic data can be labeled with behavioral context. For the purpose of this task, it is assumed that the behavior that follows the vocalization is the intended meaning of the vocalization. Two common behavioral-labeled rumbles for our subjects were contact rumbles, in which a reply from another elephant shortly followed, and "Let's Go" rumbles, which were associated with the elephants moving to a different part of the yard. This task is designed to determine whether there is acoustic variation between these two rumbles made in different behavioral contexts. As in the previous tasks, a number of experiments are performed to show the effect of various parameters.

Subjects

The same adult male and five adult females that participated in the call type experiment were used in this experiment. See Table 5.1 for the number of vocalizations used from each subject.

Data Collection

The data was collected in the same manner as in the call type experiment. All vocalizations were amplitude scaled using variance normalization to normalize their power.

Feature Extraction

Features were extracted using a 300ms Hamming window with 100ms frame step size. The same Greenwood warping constants and equal loudness curve used in the call type experiment were used in this experiment. The other feature extraction parameters will be discussed in the results section.

Model Parameters

The same DTW system used in the speaker identification experiment was used in this experiment. A three state left-to-right HMM was used to model each class with an additional three state HMM to model the silence both following and preceding the vocalization. All state observations were modeled with a single multivariate Gaussian. More details are discussed in the call type experiment section.

		Classification	
		Contact	Let's Go
Label	Contact	9	9
	Let's Go	7	14

MFCC-DTW (59.0%)

		Classification	
		Contact	Let's Go
Label	Contact	7	11
	Let's Go	6	15

MFCC-HMM (56.4%)

		Classification	
		Contact	Let's Go
Label	Contact	7	11
	Let's Go	10	11

MFCC-HMM with F B Compression (46.2%)

		Classification	
		Contact	Let's Go
Label	Contact	6	12
	Let's Go	7	14

gPLP-HMM (51.3%)

Figure 5.9 – Behavioral Context Results

RESULTS

The baseline system for this task is a MFCC-DTW system similar to the one used in the speaker identification task. HTK 3.2.1 (Cambridge University Engineering Department, 2002) was used to calculate eighteen MFCCs using 40 filter banks spaced between 10Hz and 3000Hz, the same filter bank used in the estrous cycle determination experiment. A pre-emphasis coefficient of 0.95 was used, and as in the other tasks, log energy was added as a feature. The classification matrix for this experiment is in Figure 5.9 at the upper left. The classification accuracy for the experiment is 59.0%. Although this sounds promising, chance classification accuracy on a two-class classification experiment is 50.0%, and this was one of the highest accuracies acquired with this system. As the following experiments will show, this high classification accuracy was probably an outlier.

The HMM classification model was applied to this task as well. To compare the HMM system with the DTW system, the feature extraction parameters used in the previous

experiment were used here to calculate eighteen coefficients. The classification matrix for this experiment is in Figure 5.9 at the upper right. The classification accuracy of 56.4% is also marginally above chance. Therefore, it is still questionable as to whether these two rumbles can be discriminated.

An attempt to improve classification accuracy by compressing the filter bank range and emphasizing the lower frequencies was made in the next experiment. The compressed filter bank is the same used in the estrous cycle experiment, 40 filters spaced between 10Hz and 300Hz. The confusion matrix for this experiment is in Figure 5.9 at the lower left. The classification accuracy for this experiment is 46.2%. Since this is below chance classification, this system was definitely unsuccessful. The fact that the accuracy decreased when the filter bank range was compressed to capture more of the rumble's energy is another indication that this system can not discriminate between the two rumbles.

Finally, the gPLP model is applied to the task. Eighteen gPLP coefficients were extracted using the same filter bank and parameters as the previous MFCC experiment. The classification accuracy of this experiment was 51.3% and the confusion matrix in Figure 5.9 at the lower right. This system also failed to successfully classify the rumbles by behavioral context.

Because the classification accuracies for this task are near chance, it would be hard to make the argument that the gPLP framework can successfully perform this classification task. This does not mean that the elephant subject cannot tell them apart, only that discriminating features and classification model are not being used. Currently, however, the gPLP framework consisting of the possible classification and feature extraction models discussed in chapter 4 lacks the ability to perform this classification. Additional features, a different HMM topology, or noise reduction techniques could make discrimination possible.

The final section of this chapter applies traditional statistical hypothesis tests to gPLP features for the purpose of comparison against the HMM and DTW classification models. A brief background of statistical hypothesis tests is presented in chapter 4. Although the statistical tests do not explicitly classify vocalizations, they can show which features are most useful for classification and whether the data is grouped into significantly different clusters.

Statistical Tests

While classification accuracy results show the ability of the classification system to generalize with respect to unseen data, they do not provide a measure of the relative distance between each class. To test the statistical significance of the differences between each class, multivariate analysis of variance (MANOVA) is performed on the same four African elephant experiments as in chapter 5. The analysis is performed with and without the state information as a second independent variable. In all experiments, a three state left-to-right HMM model trained on all of the data was used to align the vocalizations. No silence models were used, hence the first and third state of each HMM primarily represents the noise before and after the vocalization. Eighteen gPLP coefficients and log-energy were calculated using the compressed filter bank described in each task's section.

A summary of the MANOVA results using Wilk's Λ statistic is shown in Table 5.5 ordered from least significant separation of classes to most significant. The F value is a measure of the similarity between the different classes of data. A larger F value indicates greater separation. The first subscript on the F value indicates the degrees of freedom of the data and the second subscript indicates the number of data vectors used to calculate the F value. In this experimental context, the number of data vectors is exactly the number of frames extracted from the vocalizations. The P value, calculated from the F value and its subscripts, is the probability that the data does not fall into the labeled classes. Therefore, a

	With State Information	Without State Information
Behavioral Context	$F_{38,3723} = 7.329$, $P<0.001$	$F_{19,3723} = 17.317$, $P<0.001$
Estrous Cycle	$F_{76,9172} = 13.190$, $P<0.001$	$F_{38,9172} = 39.611$, $P<0.001$
Speaker Identification	$F_{190,13426} = 52.089$, $P<0.001$	$F_{95,13426} = 143.975$, $P<0.001$
Call Type	$F_{152,5051} = 73.337$, $P<0.001$	$F_{76,5051} = 210.081$, $P<0.001$

Table 5.5 – MANOVA Results

small P value indicates that the data is correctly labeled and can be separated into the labeled classes. In a typical hypothesis test, probabilities less than 0.05 are usually considered low enough to invalidate the hypothesis which is made in the negative sense such as "The data cannot be separated into the labeled classes."

The results are consistent with the classification experiments when the relative difficulty of each task is compared. The task with the worst classification rate, the behavior context task, had the least significant class separation in MANOVA analysis, while the call type task, the task with the best classification rate (on the clean data), had the most significant class separation. The estrous cycle and speaker identification tasks are in the appropriate order as well.

The biggest difference between the MANOVA results and the supervised classification results is that while the HMM system could not perform the behavioral context task, the MANOVA results show that the classes are separated with significant probability. Looking at the MANOVA with state information results more closely, there were six gPLP coefficients with significance probabilities greater than 0.35 which indicate that these coefficients are not significantly different between the two rumbles. These coefficients could affect the HMM system enough to make it ineffective in classifying the vocalizations. One way to remedy this problem could be a feature pre-processing step which keeps only those gPLP coefficients that contribute to the classification and discards the gPLP coefficients that may confuse the system.

Summary

This chapter discussed the results of a number of supervised classification tasks on African elephant vocalizations. A number of different classification models and feature extraction models were used to show that the gPLP framework using the gPLP feature extraction model, paired with an HMM classification model achieves comparable, if not better, accuracies when compared to standard human feature extraction models. The effects of the various modifications made to the standard human feature extraction models were also highlighted. The results were also better when compared with traditional bioacoustic features on the speaker identification task. The next chapter will explore the use of the gPLP framework for the unsupervised classification of beluga whale vocalizations.

Chapter 6

UNSUPERVISED CLASSIFICATION OF BELUGA WHALE VOCALIZATIONS

Background

BELUGA WHALES

Beluga whales are one of the most vocal of the Odontocetes (Sjare and Smith, 1986b). Belugas are a highly social species and primarily live along the coasts in north circumpolar habitats. The vocal repertoire of beluga whales has been decomposed up into a number of distinct types of vocalizations in a number of studies (Faucher, 1988; Recchia, 1994; Sjare and Smith, 1986b). Although beluga whales make a number of different types of sounds, they also generate vocalizations that fall between these different classes of vocalization types, leading some researchers to describe the repertoire as graded (Sjare and Smith, 1986b). The lack of discrete vocalization types makes behavior correlation extremely difficult. Nonetheless, Faucher (1988) and Sjare and Smith (1986a) present a few behavior correlations with their classification methods. These correlation studies are also hampered by the fact that beluga whales spend large portions of times underwater, outside the visual range of the researchers.

Sjare and Smith (1986b) and Faucher (1988) divide wild beluga whale vocalizations into four types: whistles, clicks, pulsed tones, and noisy calls. The Sjare and Smith (1986b) study used data collected from beluga whales in Cunningham Inlet, Northwest Territory, Canada, while the Faucher (1988) study collected vocalizations from beluga whales in the Saint Lawrence River. Both studies focused on the whistles and divided the whistles into seven contour types, most with multiple subtypes. Both studies note that the whistles in each

category are extremely variable, and the repertoire should be described as graded even though all whistles could be placed in one of the seven contour types.

Recchia (1994) used a slightly different classification scheme and divided captive beluga vocalizations into eight main types, some with subtypes, based on spectrogram analysis. The main types were Clicks, Jaw Claps, Yelps, Chirps, Whistles, Trills, Buzzsaws, and Screams. Chirps and Whistles, because they were the most common, were subdivided into subtypes which include Noisy Chirps, Chirp Combinations, Shifting Whistles, Noisy Whistles, and Whistle Combinations. Recchia (1994) validated the classification using AcouStat (Fristrup and Watkins, 1992) to generate acoustic features. Linear discriminant function analysis, principal components analysis (PCA), and a tree-based classifier were then used to show that the vocalizations could be divided into the vocalization types defined using acoustic features.

Recent research has shown that beluga whales exhibit the Lombard effect (Scheifele, 2003), which occurs when a speaker increases the amplitude of vocalizations in response to environmental noise. This effect was first discovered in humans (Lombard, 1911), but has since been found in a number of species (Egnor *et al.*, 2003; Manabe *et al.*, 1998; Nonaka *et al.*, 1997; Potash, 1972; Sinott *et al.*, 1975). The classification techniques discussed here were used in Scheifele (2003) to group similar vocalizations together and determine how their amplitudes varied in the presence of boat traffic. To better understand the classification techniques used, a brief background of unsupervised classification techniques will be given before the results of the experiments.

Unsupervised Classification

Unsupervised classification has a different purpose than the supervised classification tasks discussed in the previous chapter. Unsupervised classification is a machine learning technique which finds the natural groupings of the data as opposed to supervised

classification techniques which classifies data into predetermined groups. It is closely related to clustering techniques. In many cases, an experiment can be described in either framework. The main difference is that unsupervised classification has the intent of creating a classification system, while clustering is more concerned with the determination and isualization of the natural groupings. It is most commonly used to create a classification system when class labels are not present.

One common unsupervised classification technique used here is a competitive neural network (Hagan *et al.*, 1996). It is a generalization of the basic K-means algorithm. Each cluster of data is not represented by the mean of the cluster, but instead by a neuron, which is much more flexible than a mean vector as in traditional K-means since it can incorporate a wide variety of similarity scores and models. A competitive neural network is much different from the well-known feed-forward artificial neural network because the nodes are unconnected. Each node stands alone and represents the center of each natural cluster of the data.

The general competitive neural network algorithm is shown below:

1. Initialize starting parameters of the neurons
2. Classify each data point by choosing the nearest neuron
3. With the labels obtained in step 2, recalculate the neuron parameters to center the neuron in the cluster
4. Iterate steps 2-3 until the neuron parameters converge

Initialization is accomplished by either random initialization in the feature space, or by slightly perturbing the center point of the entire dataset differently for each neuron. The number of neurons in the network is chosen a priori. There are techniques for estimating the number of classes in the dataset, but they are not used here. Iterations are performed until the neuron parameters are relatively stable.

The novel component of the competitive neural network used in this study is that each neuron is represented by an HMM. The maximum likelihood of each vocalization being generated by the HMM as determined by the Viterbi algorithm (Forney, 1973) is used as the distance metric in step 2. The Baum-Welch algorithm (Baum, 1972; Baum *et al.*, 1970; Moon, 1996) is used to determine the new HMM parameters in each iteration in step 3. Therefore, each neuron represents the maximum likelihood HMM model for each cluster. This representation allows the neurons to model temporal and spectral differences between each type of beluga whale vocalization.

The unsupervised classification system discussed above, incorporating the gPLP model to extract features and the HMM-based competitive neural network for classification, will be used to classify beluga whale vocalizations into clusters of similar vocalizations. The results of the classification will then be qualitatively analyzing and compared to expert labeling of the data. Finally, the elephant call type vocalization data from the previous chapter will be clustered using the unsupervised classification technique to validate the model.

Subjects

The vocalizations were made by a small, endangered population of about 700 beluga whales residing in the Saint Lawrence River Estuary year round (Scheifele, 2003). The population is geographically isolated from other populations of its species (Kingsley, 1998). This population is unique because it is sub-arctic and has been studied extensively over the last twenty years. Studies of the population include seasonal distribution, size, age structure, toxicology and pathology (Beland *et al.*, 1993; Kingsley, 1998; Martineau *et al.*, 1994). It is believed that this population is especially susceptible to the effects of anthropogenic noise due to the large amount of shipping traffic in the Saint Lawrence River.

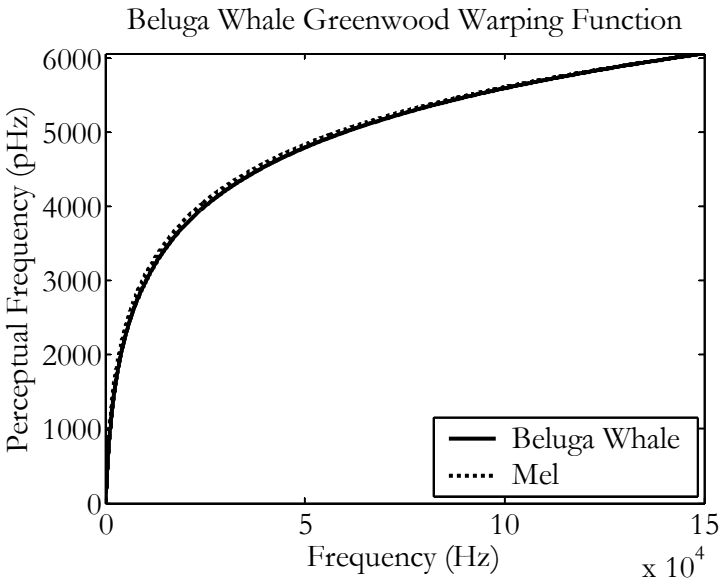


Figure 6.1 – Beluga Whale Greenwood Warping Function

Data Collection

The vocalizations were collected at four different sites near the intersection of the Saguenay River and the St. Lawrence Estuary. The beluga whales were recorded in July and August at 7:00 a.m., 10:00 a.m., and 2:00 p.m. for one hour at a time. A total of 230 hours of recordings were made over the course of 6 years by Schiefele (2003). The acoustic data was collected with an omni directional hydrophone (International Transducer Corporation Model ITC-1042) with pre-amplifier and recorded on a Sony TCD-D8 DAT tape with 16-bit quantization. A number of the vocalizations were isolated using PRAAT 4.1 (Boersma and Weenink, 2003) and saved as individual files, which were then used in this analysis. A total of 67 vocalizations of various types were available for this experiment. Additional details about the recording methodology can be found in Scheifele (2003).

Feature Extraction

Features were extracted using a 30ms window with a 10ms step size. Eighteen gPLP coefficients and log energy were used as features to create a 19 element feature vector. The

Frequency (Hz)	100	300	1000	3000	10000	30000	100000	125000
Hearing Threshold (dB re 1μP)	127.0	118.0	112.0	84.0	62.0	40.0	64.0	118

Table 6.1 – Beluga Whale Audiogram Data

Greenwood warping function constants were calculated using an approximate hearing range of 100Hz to 150kHz as determined by Scheifele (2003). The Greenwood warping function constants were calculated using the equations from chapter 4 as follows:

$$k = 0.88, \tag{6.1}$$

$$A = \frac{F_{\min}}{1-k} = \frac{100}{1-0.88} = 833.333, \text{ and} \tag{6.2}$$

$$a = \log_{10}\left(\frac{F_{\max}}{A} + k\right) = \log_{10}\left(\frac{150000}{833.333} + 0.88\right) = 2.257. \tag{6.3}$$

A graph of the beluga whale warping function compared to the Mel scale is in Figure 6.1. In this figure, the Mel scale has been normalized to the same range as the beluga whale Greenwood warping curve for the sake of comparison. Although the hearing range of the beluga is much larger than that of a human, the Greenwood warp for a beluga whale is very similar to the Mel scale. Since 100Hz is used as $F_{\min}$, the lower bound of the filter bank must be greater than 100Hz, otherwise negative values of f_p will result. This constraint is explained further in chapter 4.

The equal loudness curve is calculated using audiogram data from Ketten (1998) as outlined in chapter 4. The equal loudness curve, using the data in Table 6.1 and the polyfit function in MATLAB®, is shown in Figure 6.2. The equation for the curve shown is

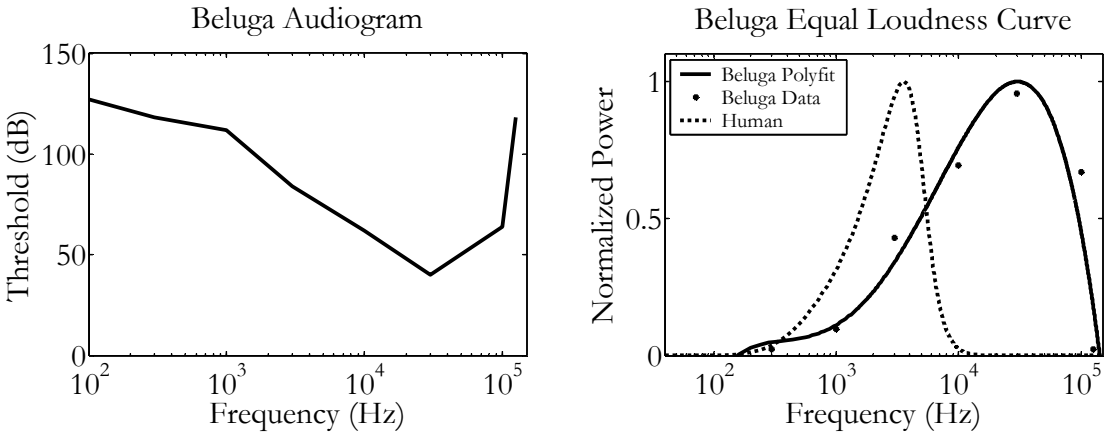


Figure 6.2 – Beluga Whale Audiogram and Equal Loudness Curve

$$\hat{E}(f) = \begin{array}{c} -5.8561\times10^{-1} f^4 + 1.7285\times10^{1} f^3 - 1.8393\times10^{2} f^2 + \\ 8.4856\times10^{2} f - 1.4397\times10^{3} \end{array}. \qquad 6.4$$

The larger range of hearing of the beluga whale is apparent in the plot by the much wider equal loudness curve. The high frequency slope is about the same between the human and beluga whale equal loudness curves, but the low frequency side of the curve has a much more gradual decline in the beluga whale curve. The evolutionary basis for this adaptation is largely unknown.

The filter bank used for analysis was spaced between 10Hz and 8000Hz and had 60 filters. This range was chosen to make the filter bank for all vocalizations the same. Since the sampling rate of the vocalizations varied from 16kHz to 44kHz, choosing the Nyquist frequency as the upper limit of the filter bank would have resulted in different filter banks depending on the sampling rate of the vocalization. The number of filters was chosen to be near the maximum number allowed to calculate accurate filter energies for the lower frequency filters as described in chapter 4. Using 60 filters in the filter bank, the bandwidth of the filters is approximately 140Hz at 1kHz. Although there is no widely available critical bandwidth data available for the beluga whale, critical ratio data (Johnson *et al.*, 1989) infers

that beluga whales have a critical bandwidth near 25Hz at 1kHz which is consistent with other aquatic mammals.

Model Parameters

A five state left-to-right HMM was used for each neuron model. This topology was chosen because while some vocalizations are short, there is a large amount of frequency modulation in a number of the vocalizations. The silence before and after the vocalizations was not modeled separately and was included in the five state neuron model. All state observations were modeled with a single multivariate Gaussian. The Hidden Markov Model Toolkit (HTK) version 3.2.1 from Cambridge University (2002) was used, with modification, to implement the feature extraction and HMM functionality. The experiment was performed a number of times using neural networks with five and ten neurons. Two different networks were used to compare the results for a more complete analysis of the classification.

Results

The results from five separate runs of the unsupervised clustering algorithm using five clusters are shown in Table 6.2. A perfect classification would assign each type of vocalization to a different cluster. Each letter a-e designates the cluster center that each vocalization is closest to at the end of 25 iterations. The vocalizations were listed to place vocalizations with similar expert labels (in the filenames) near each other. The table of data shows that the clusters varied to a significant extent between each run.

Filename	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
*/voca1.lab	a	a	a	e	c
*/voca2.lab	a	a	a	e	c
*/vocb1.lab	a	a	a	e	c
*/DWNWARBLE1BSMCASS6.lab	a	d	b	b	c
*/DWNWARBLE2BSMCASS6.lab	a	d	e	d	c
*/DWNWARBLE3BSMCASS6.lab	a	d	b	b	c
*/QUADBUZZA1BSMCASS6a.lab	a	d	d	b	c
*/QUADBUZZA1BSMCASS6b.lab	a	d	d	b	c
*/QUADBUZZA1BSMCASS6c.lab	a	d	d	b	c
*/QUADBUZZA1BSMCASS6d.lab	a	d	d	b	c
*/BUZZA1BSMCASS6.lab	a	d	d	b	c
*/SAG6-01T2CHIRPA.lab	a	b	c	d	a
*/SAG6-01T3CHIRPA.lab	a	b	c	d	a
*/CHIRPA28.lab	a	b	c	b	a
*/CHIRPA31MOS8_9_01.lab	a	b	c	d	a
*/CHIRPA32MOS8_9_01.lab	a	b	c	d	a
*/CHIRPA33MOS8_9_01.lab	a	b	c	d	a
*/CHIRPA34MOS8_9_01.lab	a	b	c	d	a
*/CH6-01T5CHIRPA.lab	c	a	b	e	d
*/CH6-01T6CHIRPA.lab	c	a	b	e	d
*/CH6-01T7CHIRPA.lab	c	a	b	e	d
*/CH6-01T8CHIRPA.lab	c	a	b	e	d
*/CH06-01T1CHIRPA.lab	c	e	e	e	a
*/CH6-01T2CHIRPA.lab	c	e	e	e	a
*/CH6-01T4CHIRPB.lab	c	c	b	e	d
*/SAG6-01T10CHIRPB.lab	a	b	c	d	a
*/SAG6-01T12CHIRPB.lab	a	b	c	d	a
*/SAG6-01T13CHIRPB.lab	a	b	c	d	a
*/SAG6-01T14CHIRPB.lab	a	b	c	d	a
*/SAG6-01T15CHIRPB.lab	a	b	c	b	a
*/SAG6-01T16CHIRPB.lab	a	b	c	d	a
*/SAG6-01T4CHIRPB.lab	a	b	c	d	a
*/SAG6-01T5CHIRPB.lab	a	b	c	d	a
*/SAG6-01T6CHIRPB.lab	a	b	c	d	a
*/SAG6-01T7CHIRPB.lab	a	b	c	b	a
*/SAG6-01T8CHIRPB.lab	a	b	c	d	a
*/SAG6-01T9CHIRPB.lab	a	b	c	d	a
*/whisa1.lab	b	d	b	b	a
*/WHISA2.lab	e	d	b	b	a
*/whisa3.lab	e	d	b	b	a
*/whisb1.lab	a	e	a	b	e
*/whisc1.lab	c	e	b	e	e
*/whisc2.lab	c	e	b	e	e
*/whisc3.lab	c	e	b	e	c
*/whisd1.lab	c	e	e	e	a
*/whisd2.lab	c	e	e	a	a
*/whisd3.lab	c	e	e	a	a
*/whise1.lab	b	d	d	e	c
*/whise2.lab	b	a	a	e	d
*/DWNWHISA1.lab	b	c	d	a	d
*/DWNWHISA4.lab	b	a	d	a	d
*/DWNWHISA6.lab	b	c	d	a	d
*/DWNWHISA8.lab	b	c	e	a	d
*/DWNWHISA5.lab	e	e	e	a	d
*/DWNWHISA10.lab	e	c	d	a	d
*/DWNWHISA2.lab	e	c	d	a	d
*/DWNWHISA3.lab	e	c	d	a	d
*/DWNWHISA7.lab	e	c	d	a	d

Table 6.2 – Results from 5 Cluster Unsupervised Classification

The results of the runs are inconsistent. Groups of vocalizations are assigned different letters each experimental run and the vocalization groupings are not consistent neither. However, despite the inconsistency between runs, trends in the classification can be seen. For instance, down warbles and buzzes consistently cluster together as do a-chirps and b-chirps. This could indicate that these vocalizations might from one repertoire type. However, this clustering could also be due to the fact that not enough clusters were used in the classification.

Another trend in the data is that there is a group of chirps with filenames starting with "CH6" that consistently forms a group separate from other chirps. Although these vocalizations could be mislabeled, it is more likely that the recording conditions were different for these vocalizations, and the classification algorithm is using these recording channel differences as the basis for classification.

The classification of the whistles is the most difficult to analyze because it is the most inconsistent. This inconsistency could be due to the fact that these are the longest vocalizations with the most frequency modulation. C-whistles and d-whistles seem to have the most in common since they match up in the same cluster twice. Both c-whistles and d-whistles match up with the a-whistles once, but not at the same time. Although a-whistles, c-whistles, and d-whistles seem similar to some extent, they are probably discrete call types. However, the e-whistles never classify with the other whistles, indicating that they are probably a unique type of whistle. Interestingly, the c-whistles and d-whistles are clustered with the CH6-chirps quite often. This could mean that these vocalizations share similar recording conditions.

Filename	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
*/voca1.lab	a	i	j	a	i
*/voca2.lab	a	i	j	e	i
*/vocb1.lab	a	i	j	e	i
*/DWNWARBLE1BSMCASS6.lab	g	d	b	h	a
*/DWNWARBLE2BSMCASS6.lab	j	e	f	f	f
*/DWNWARBLE3BSMCASS6.lab	j	d	b	h	f
*/QUADBUZZA1BSMCASS6a.lab	j	e	b	h	f
*/QUADBUZZA1BSMCASS6b.lab	b	e	b	h	b
*/QUADBUZZA1BSMCASS6c.lab	e	e	a	h	b
*/QUADBUZZA1BSMCASS6d.lab	j	e	b	h	b
*/BUZZA1BSMCASS6.lab	e	e	b	h	d
*/CH06-01T1CHIRPA.lab	h	i	h	e	i
*/CH6-01T2CHIRPA.lab	h	i	h	e	i
*/CH6-01T5CHIRPA.lab	h	a	e	b	d
*/CH6-01T6CHIRPA.lab	h	a	e	b	d
*/CH6-01T7CHIRPA.lab	h	a	e	b	e
*/CH6-01T8CHIRPA.lab	h	a	e	b	e
*/CH6-01T4CHIRPB.lab	h	a	e	b	d
*/CHIRPA28.lab	b	f	d	d	b
*/CHIRPA31MOS8_9_01.lab	j	f	d	d	b
*/CHIRPA32MOS8_9_01.lab	j	d	b	d	b
*/CHIRPA33MOS8_9_01.lab	g	e	f	d	b
*/CHIRPA34MOS8_9_01.lab	j	f	a	d	b
*/SAG6-01T2CHIRPA.lab	j	f	d	d	b
*/SAG6-01T3CHIRPA.lab	g	d	d	d	b
*/SAG6-01T10CHIRPB.lab	j	d	b	d	b
*/SAG6-01T12CHIRPB.lab	g	e	f	d	b
*/SAG6-01T13CHIRPB.lab	g	e	f	d	b
*/SAG6-01T14CHIRPB.lab	j	d	a	d	b
*/SAG6-01T15CHIRPB.lab	b	f	j	d	b
*/SAG6-01T16CHIRPB.lab	g	e	d	d	b
*/SAG6-01T4CHIRPB.lab	g	e	d	d	b
*/SAG6-01T5CHIRPB.lab	g	f	d	d	b
*/SAG6-01T6CHIRPB.lab	g	f	d	d	b
*/SAG6-01T7CHIRPB.lab	g	f	j	d	b
*/SAG6-01T8CHIRPB.lab	g	e	d	d	b
*/SAG6-01T9CHIRPB.lab	j	f	d	d	b
*/whisa1.lab	b	f	j	j	f
*/WHISA2.lab	g	f	j	d	f
*/whisa3.lab	g	f	j	d	f
*/whisb1.lab	e	g	i	a	i
*/whisc1.lab	a	g	h	j	a
*/whisc2.lab	a	g	h	j	a
*/whisc3.lab	a	g	j	j	i
*/whisd1.lab	h	i	c	a	g
*/whisd2.lab	h	i	c	j	i
*/whisd3.lab	a	i	e	j	g
*/whise1.lab	d	b	d	d	e
*/whise2.lab	d	b	d	d	c
*/DWNWHISA1.lab	d	b	c	g	a
*/DWNWHISA10.lab	d	b	c	g	g
*/DWNWHISA2.lab	d	b	c	g	c
*/DWNWHISA3.lab	d	b	c	g	g
*/DWNWHISA4.lab	d	b	c	f	g
*/DWNWHISA5.lab	d	b	h	g	g
*/DWNWHISA6.lab	d	b	c	g	g
*/DWNWHISA7.lab	d	b	i	f	g
*/DWNWHISA8.lab	d	b	i	g	c
*/DWNWHISA9.lab	d	b	c	g	c

Table 6.3 – Results from 10 Cluster Unsupervised Classification

The above analysis points out a number of problems with unsupervised classification techniques. One problem is that it is difficult to explain in a qualified manner the criteria used to create the clusters. From a quantitative standpoint, maximum likelihood distance is used to create the clusters, but no simple qualitative measure is available. For this reason, it is difficult to tell whether recording channel characteristics, background noise, or individual speaker differences are significant factors in the determination of the clusters. In this experiment, all of these factors would be considered noise and we prefer these factors to have a negligible contribution to the classification.

Another problem is that the number of clusters chosen for the classification can have a large effect on the clarity of the results. Using too many clusters causes vocalizations that should be clustered together to be put in separate clusters. Too many clusters also reduces the population of each cluster, making the results less reliable and robust. The recommended population of each cluster is difficult to quantify, but 10 vocalizations for each cluster is probably near the minimum population desired. On the other hand, too few clusters causes unlike vocalizations to be clustered together. This occurs with the five cluster experiments and is apparent from the discussion.

The results from five experimental runs of the unsupervised classification algorithm using ten clusters are in Table 6.3. Many of the same trends are present in this set of classifications as in the five cluster classifications. Both kinds of chirps consistently cluster together, sometimes in more than one cluster. This strengthens the hypothesis made previously that the two chirps may be the same type of vocalization, not two discrete types. The buzzes and down warbles that clustered together in the five cluster classifications appear to cluster more separately when ten clusters are used. However, the common confusion

between buzzes and down warbles indicates that these two vocalizations are probably very similar.

The CH6-chirps form their own cluster separate from the other chirps as in the five cluster experiments. The d-whistles also cluster with the CH6-chirps as before. This supports the claim made before that these vocalizations were either recorded in similar environments or that these vocalizations are mislabeled. In contrast with the five cluster classification, the different types of whistles cluster separately more consistently, supporting the expert labeling scheme.

With the exception of a few vocalizations, the unsupervised clustering grouped vocalizations in a similar fashion as the expert (Scheifele, 2004). However, these exceptions could be due to background noise or recording conditions. Unfortunately, the dataset contained too few examples of some of the vocalizations to perform reliable analysis.

Validation of Algorithm Using Elephant Vocalizations

While the beluga whale data clustered well, it did not show whether the gPLP framework can be generalized for different unsupervised classification tasks. To show that the gPLP framework can be adapted for a different species, the elephant call type data from the previous chapter will be classified with the same unsupervised classifier. As in the previous chapter, both the original, complete dataset as well as the clean dataset will be classified. The feature extraction parameters are the same as before, 18 gPLP coefficients with 30 filters in the filter bank confined to the range 10Hz – 1500Hz. The frame size is 60ms with 20ms steps. A total of 10 clusters was used to classify the elephant vocalizations. The results from five trials of the unsupervised classification on the original, complete dataset are shown in Table 6.4, and the results from five trials on the clean dataset are shown in Table 6.5.

Filename	Call Type	Elephant	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
*/NR_1932BA.lab	N. Rumble	Bala	a	b	c	i	j
*/NR_1935BA.lab	N. Rumble	Bala	a	b	g	j	i
*/NR_D3557BA.lab	N. Rumble	Bala	a	b	c	j	i
*/NR_D5551BA.lab	N. Rumble	Bala	a	b	a	i	j
*/NR_D3731BA.lab	N. Rumble	Bala	a	j	c	i	i
*/RV_D2803BA1.lab	Rev	Bala	a	c	h	i	g
*/RV_D2803BA2.lab	Rev	Bala	a	c	h	e	i
*/CR_D914FA.lab	Croak	Fiki	a	h	c	g	a
*/NR_D434FA.lab	N. Rumble	Fiki	b	b	i	a	g
*/RV_D1318CA1.lab	Rev	Mackie	b	b	d	d	e
*/RV_D1318CA2.lab	Rev	Mackie	b	b	d	c	e
*/SN_D1132CA1.lab	Snort	Mackie	b	b	j	g	h
*/SN_D1132CA2.lab	Snort	Mackie	b	b	d	d	e
*/SN_D1132CA3.lab	Snort	Mackie	b	b	j	d	h
*/SN_D2418CA1.lab	Snort	Mackie	b	b	d	c	h
*/SN_D2418CA2.lab	Snort	Mackie	b	b	d	c	h
*/SN_D2509CA.lab	Snort	Mackie	b	b	j	g	h
*/CR_D5500CA1.lab	Croak	Mackie	b	j	h	j	b
*/CR_D5500CA2.lab	Croak	Mackie	b	j	h	j	b
*/CR_D5527CA1.lab	Croak	Mackie	b	j	h	j	b
*/CR_D5527CA2.lab	Croak	Mackie	b	j	h	j	b
*/TR_2117MA.lab	Trumpet	Moyo	b	b	b	c	b
*/TR_D5503MA.lab	Trumpet	Moyo	b	f	a	h	a
*/SN_D3712FA1.lab	Snort	Fiki	b	f	d	c	a
*/SN_D3956BA.lab	Snort	Bala	b	g	a	g	h
*/NR_D10039BA.lab	N. Rumble	Bala	b	h	d	j	b
*/TR_D3217RA1.lab	Trumpet	Robin	c	e	j	i	a
*/CR_D5144CA1.lab	Croak	Mackie	c	f	e	c	f
*/CR_D5144CA2.lab	Croak	Mackie	c	f	e	c	f
*/CR_4603BA.lab	Croak	Bala	c	i	d	a	f
*/RV_D4145BAx.lab	Rev	Bala	d	c	j	e	j
*/RV_D953BAx.lab	Rev	Bala	d	c	a	f	e
*/CR_D2513FA.lab	Croak	Fiki	d	d	b	i	e
*/CR_D2624FA.lab	Croak	Fiki	d	d	b	i	e
*/CR_D4620CA1.lab	Croak	Mackie	d	d	i	g	j
*/CR_D4620CA2.lab	Croak	Mackie	d	d	a	f	j
*/CR_D2922CA1.lab	Croak	Mackie	d	i	b	i	g
*/TR_10110FA.lab	Trumpet	Fiki	d	f	b	e	c
*/TR_2056FA2.lab	Trumpet	Fiki	d	f	b	e	a
*/TR_D3944TA.lab	Trumpet	Thandi	d	f	j	e	a
*/NR_5625BA.lab	N. Rumble	Bala	e	b	g	a	b
*/RV_2920BAx.lab	Rev	Bala	e	c	j	e	e
*/RV_D1611BAx.lab	Rev	Bala	e	c	j	f	e
*/TR_1632MA.lab	Trumpet	Moyo	e	c	c	j	d
*/CR_D3424CA.lab	Croak	Mackie	e	d	i	h	j
*/CR_D1402CA.lab	Croak	Mackie	e	h	d	d	j
*/SN_1048RA.lab	Snort	Robin	e	d	i	a	h
*/SN_D5139TA.lab	Snort	Thandi	e	d	i	a	j
*/SN_D5644RA.lab	Snort	Robin	e	e	j	i	g
*/RV_D1311BAx.lab	Rev	Bala	f	c	h	j	c
*/RV_3412BAx.lab	Rev	Bala	f	d	c	e	e
*/TR_D3217RA2.lab	Trumpet	Robin	f	c	j	h	a
*/TR_2056FA1.lab	Trumpet	Fiki	f	f	j	c	c
*/SN_D1856BA.lab	Snort	Bala	g	a	d	i	i
*/RV_D2458BAx.lab	Rev	Bala	g	c	a	f	e
*/RV_5728BAx.lab	Rev	Bala	g	d	a	f	e
*/RV_D4643BAx.lab	Rev	Bala	g	h	d	a	e
*/TR_4507RA.lab	Trumpet	Robin	g	f	j	i	e
*/TR_D1743RA.lab	Trumpet	Robin	g	b	b	g	a
*/NR_D2458BA.lab	N. Rumble	Bala	h	c	a	j	b
*/NR_D5506BA.lab	N. Rumble	Bala	h	c	g	j	b
*/NR_D5536BA.lab	N. Rumble	Bala	h	j	g	j	j
*/SN_D1607FA.lab	Snort	Fiki	h	d	h	g	c
*/TR_753MA.lab	Trumpet	Moyo	h	c	g	h	b
*/TR_1704M.lab	Trumpet	Moyo	h	d	g	h	b
*/TR_1348MA.lab	Trumpet	Moyo	h	h	e	h	b
*/TR_D10215BA.lab	Trumpet	Bala	i	c	a	f	e
*/RV_D1025TA.lab	Rev	Thandi	i	d	d	d	g
*/SN_D356TA.lab	Snort	Thandi	i	d	d	f	a
*/SN_D3704BA.lab	Snort	Bala	i	d	d	c	c
*/SN_D3712FA2.lab	Snort	Fiki	i	f	d	f	i
*/SN_D3712FA3.lab	Snort	Fiki	i	f	d	g	a
*/CR_D2922CA2.lab	Croak	Mackie	i	i	b	i	g
*/CR_D2922CA3.lab	Croak	Mackie	i	i	b	i	j

Table 6.4 – Results from 10 Cluster Original Elephant Call Type Data

The trends in this classification task are much more difficult to spot than in the beluga whale task. Overall, however, it can be seen that the unsupervised classification system is attempting to classify each call type from each speaker into its own cluster. This trend can be seen in both the original and clean dataset. For instance, cluster a in the first trial of the original dataset contains most of Bala's noisy rumbles. Cluster b contains most of Mackie's vocalizations, but because he had a number of croaks in the dataset, the croaks were spread between three clusters. In the clean dataset, the clusters are more apparent with Mackie's croaks generally forming one cluster, and Bala's snorts forming another. There are a number of other similar clusters in the clean data results.

The main difference between the two datasets is that the clean dataset clusters more consistently. The main reason for this is the cleanliness of the data. Similar background noise between vocalizations can make two vocalizations which are not similar classify together. Neither of these datasets, however, cluster as consistently as the beluga data. The reason for this is that the natural clusters of the elephant data seem to be related to both the speaker and call type as opposed to just the call type as in the beluga data. Because this would generate 30 total clusters (5 types x 6 speakers) in the elephant data, there are not enough vocalizations in either dataset to adequately fill each cluster. Even 10 clusters is too many clusters to use for the clean data since this amounts to only three vocalizations on average for each cluster. Ten vocalizations in each cluster is usually considered the minimum. This lack of data results in inconsistent clustering since it forces multiple clusters into one.

Filename	Call Type	Elephant	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
*/CR_D1402CA.lab	Croak	Mackie	a	g	e	c	i
*/CR_D3424CA.lab	Croak	Mackie	a	g	e	c	c
*/TR_10110FA.lab	Trumpet	Fiki	c	d	h	a	e
*/TR_2056FA2.lab	Trumpet	Fiki	c	d	h	a	j
*/TR_1348MA.lab	Trumpet	Moyo	c	g	j	d	f
*/TR_753MA.lab	Trumpet	Moyo	c	h	j	g	a
*/CR_D2624FA.lab	Croak	Fiki	d	j	d	a	j
*/NR_1935BA.lab	N. Rumble	Bala	d	b	e	a	d
*/NR_D434FA.lab	N. Rumble	Fiki	d	b	j	a	f
*/TR_D5503MA.lab	Trumpet	Moyo	d	a	f	h	j
*/SN_D5644RA.lab	Snort	Robin	e	a	g	b	h
*/NR_1932BA.lab	N. Rumble	Bala	f	b	h	f	f
*/NR_5625BA.lab	N. Rumble	Bala	f	g	h	a	i
*/NR_D5551BA.lab	N. Rumble	Bala	f	b	h	a	e
*/TR_4507RA.lab	Trumpet	Robin	f	d	b	g	a
*/CR_4603BA.lab	Croak	Bala	g	d	e	d	i
*/NR_D5536BA.lab	N. Rumble	Bala	g	g	b	g	c
*/RV_D1311BAx.lab	Rev	Bala	g	g	g	h	d
*/RV_2920BAx.lab	Rev	Bala	h	j	b	g	a
*/RV_5728BAx.lab	Rev	Bala	h	j	i	d	g
*/RV_D1611BAx.lab	Rev	Bala	h	j	b	g	g
*/RV_D4145BAx.lab	Rev	Bala	h	j	i	g	g
*/RV_D953BAx.lab	Rev	Bala	h	j	i	g	g
*/TR_2056FA1.lab	Trumpet	Fiki	h	e	a	j	a
*/TR_2117MA.lab	Trumpet	Moyo	h	g	d	h	f
*/TR_D1743RA.lab	Trumpet	Robin	h	e	g	a	e
*/TR_D3217RA1.lab	Trumpet	Robin	h	a	b	j	g
*/TR_D3217RA2.lab	Trumpet	Robin	h	a	a	b	c
*/RV_D2458BAx.lab	Rev	Bala	i	d	b	f	j
*/SN_D1856BA.lab	Snort	Bala	i	c	a	f	d
*/SN_D3704BA.lab	Snort	Bala	i	i	a	f	d
*/CR_D2513FA.lab	Croak	Fiki	j	j	a	a	j
*/NR_D2458BA.lab	N. Rumble	Bala	j	d	d	f	j
*/TR_1704M.lab	Trumpet	Moyo	j	h	d	j	f
*/TR_D10215BA.lab	Trumpet	Bala	j	j	a	g	a

Table 6.5 – Results from 10 Cluster Clean Elephant Call Type Data

Summary

This chapter showed how the gPLP framework consisting of the gPLP feature extraction model and the HMM classification model can be used to perform an unsupervised classification task. The results of the classification using these features were consistent with expert labels and revealed that some established call types may be more closely related then previously thought, while other established vocalization labels may encompass more than

one vocalization type. An unsupervised classification was also performed on the elephant vocalization data from the previous chapter to validate the model.

Chapter 7

CONCLUSION

Results Summary

The experiments presented in this dissertation show the applicability of gPLP framework to bioacoustic signal analysis and classification. In supervised classification tasks on African elephant vocalizations, the best classification accuracies on the call type, speaker identification, estrous cycle, and behavioral context experiments using the gPLP model and HMM were 91.4%, 85.3%, 74.5%, and 51.3%, respectively. The results from the MANOVA statistical analysis were mostly consistent with the supervised classification results and showed a statistical difference between the classes in all experiments. The unsupervised classification experiment produced clustering results consistent with human categorization of the vocalization types using five and ten clusters.

Analysis

There are a number of factors that affect classification accuracies. The most important of these is the quality of the data. This is a major problem since unmatched noise conditions in training and testing data can greatly reduce classification accuracy. The elephant call type classification experiment shows the effect of noise on the gPLP framework. The classification accuracy increased from 79.7% to 91.4% when poor quality vocalizations were removed. The removal of poor quality vocalizations is standard practice in bioacoustic studies where the data is typically separated into different classes of quality. Once the vocalizations are separated by quality, the analysis is only performed on the better quality vocalizations.

Although the accuracy decreased when poor quality vocalizations were included, the degradation was not catastrophic. This shows that the gPLP framework is robust to noise, at least to some degree. If the framework were not robust, the classification accuracy would have dropped off significantly with the introduction of a number of noisy vocalizations. The features and model both contribute to this robustness. This is important for application to bioacoustic signals, since the recording environments are hard to control, and noisy data is very common.

Another factor that can lead to lower classification accuracies is the lack of knowledge about many species' vocal production and sound perception systems. This information can be used to improve the accuracy of the classification system using the gPLP model which can incorporate ERB or audiogram data. However, without this information, the filter bank cannot be designed properly and may emphasize the wrong portions of the spectrum or contain too many filters.

Finally, it is very difficult to test the validity of the labels for the training data. Because we lack a clear understanding of what the animal is trying to communicate most of the time, the wrong behavioral cues could be used to label a vocalization. Animals typically use the same general type of vocalization in different behavioral contexts. For example, elephants appear to use rumbles to communicate a number of things (Poole *et al.*, 1988). This especially affects the estrous cycle determination task which probably includes rumbles not meant for reproductive purposes. The rumbles with other meanings occurring in the particular estrous phase introduce noise into the model.

Applicability of gPLP Framework

Although African elephants and beluga whales were used as examples in this study, this framework is easily applicable to other species. Species-specific frequency warping functions

and equal loudness curves can be incorporated into the gPLP feature extraction model to account for each species' unique auditory perception abilities. The rate of change of the vocalizations dynamics can be modeled by adjusting the analysis window and window step size. The HMM can be lengthened or shortened to account for different degrees of complexity in the vocalization. For those species which use a syllabic structure, a language model can be applied to the system to model the grammar of the vocalization. These changes encompass much of the variation that occurs between species.

Contributions

The dissertation introduces the gPLP framework for the analysis and classification of animal vocalizations which can be generalized to any number of species. In particular, the gPLP feature extraction model incorporates information about the perceptual abilities of the species being studied. Existing data from common experimental data such as audiograms and equal rectangular bandwidth tests can be incorporated into the feature extraction model to create a perceptual spectrogram that more closely resembles what is heard as opposed to the generic spectrogram. The gPLP framework is applied to African elephant and beluga whale vocalizations to demonstrate its applicability. It is also used to perform a number of different tasks including supervised classification systems, statistical hypothesis testing, and unsupervised classification systems.

Future Work

There are a number of contributions and advancements to be made on the gPLP framework. First, because of the small critical bands of many animal species, filter bank design can be problematic since the frequency resolution of the spectrum is not fine enough at the lower frequencies. One possible solution to this would be to use a different spectral

estimation method besides the FFT. Wavelets might be one possible solution since the analysis window increases for the lower frequencies, resulting in improved resolution.

The data collected for bioacoustic tasks is often noisy due to the desire to record the animals in a naturalistic environment. Therefore, signal enhancement techniques that enhance the vocalization while suppressing background noise would be extremely helpful. The effect of noisy data can be seen in the African elephant call type classification experiment. Once the noisy data was removed, there was a significant improvement in the classification accuracy.

Although the gPLP framework provides a feature extraction model, there are many other possible features that could also be incorporated. The gPLP feature extraction model heavily borrows from speech processing research, but bioacoustics research typically uses completely different types of features for analysis. The addition of algorithms that can extract these typical bioacoustics features for use in the classification models could improve classification accuracies. The main issue with incorporating traditional bioacoustics features is that they are typically extracted on a vocalization-basis, not a frame-basis. This creates a temporal resolution mismatch between the traditional bioacoustic features and the gPLP features, which is not easily remedied. Modifications would be needed to be made to the traditional HMM architecture to handle features with various temporal resolutions.

The unsupervised classification results, although validated by an expert in this study, could be strengthened through the use of metrics that determine the reliability of the results. Some examples of such metrics are consistency measures, the distance from each data point to its representative neuron or cluster center, the distance between the cluster centers, and the ratio between cluster density and cluster center distance. The validation of unsupervised

clustering algorithms is a new research area, therefore these methods will have to be developed and tested since there are few already existing.

As the gPLP framework is applied to other species, there will certainly be other issues that need to be addressed. For instance, many animals, such as birds, use a type of song for communication. In these cases, a language model would be beneficial for classification tasks. Another consideration for other species is the appropriate analysis window. As the rate of frequency modulation increases, the size of the analysis window should decrease to maintain signal stationarity. An automatic method which could detect the pitch and adjust the window size dynamically, known as variable window length processing in speech, would provide a way to resolve the tradeoff between analysis window size and frequency resolution.

Summary

Generalized Perceptual Linear Prediction (gPLP) analysis generates efficient, perceptually meaningful features for animal vocalizations through the incorporation of information about each species perceptual abilities. Coefficients calculated using the gPLP model represent the shape of the vocal tract filter during production, but as shown in the elephant examples, can include harmonic information as well. The gPLP feature extraction model is adaptable to any number of species as long as some basic information about their perceptual abilities is known. The coefficients are robust to noise, relatively uncorrelated, consistent in a Euclidean space, and extremely efficient.

The gPLP framework can be used for a variety of purposed tasks including visualization through perceptual spectrograms, hypothesis testing through statistical analysis, and as a feature extraction algorithm to various types of automatic classification systems. The use of these systems can speed up analysis by classifying vocalizations quickly and can reveal complex schemes in language and acoustic structures.

One possible benefit of this research is that understanding how animals communicate can lead to new technologies for human use such as echolocation and coding methods. The main benefit, however, is the potential for humans to better co-exist with the multitude of animal species present in our environment. We currently communicate on a basic level with other species using mostly body language. Understanding acoustic components of their communication could lead to better diagnosis of their desires and needs, such as the environment preferred, reproductive concerns, or medical treatment required. The better species can communicate with each other, the better they can share the resources of our planet.

BIBLIOGRAPHY

Allen, J. (**1995**). *Natural Language Understanding*. Redwood City, CA.

Anderson, S. E. (**1999**). "Speech recognition meets bird song: A comparison of statistics-based and template-based techniques," *The Journal of the Acoustical Society of America* **106**(4), 2130.

Baum, L. E. (**1972**). "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities* **3**, 1-8.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (**1970**). "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics* **41**, 164-171.

Békésy, G. V. (**1960**). *Experiments in Hearing*. New York: McGraw-Hill.

Beland, P., DeGuise, S., Girard, C., Laglace, A., Martineau, D., Michaud, R., Muir, D. C. G., Norstrom, R. J., Pelletier, E., Ray, S., and Shugart, L. R. (**1993**). "Toxic compounds and health and reproductive effects in St. Lawrence beluga whales," *Journal of Great Lakes Research* **19**, 766-775.

Berg, J. K. (**1983**). "Vocalizations and associated behaviors of the African elephant (*Loxodonta africana*) in captivity," *Z. Tierpsychol.*, 63-79.

Boersma, P., and Weenink, D. (**2003**). PRAAT: Doing phonetics by computer. The Netherlands: University of Amsterdam. Available at http://www.fon.hum.uva.nl/praat/.

Bradbury, J. W., and Vehrencamp, S. L. (**1998**). *Animal Communication*. Sunderland, MA: Sinauer Associates.

Buck, J. R., and Tyack, P. L. (**1993**). "A quantitative measure of similarity for *tursiops truncatus* signature whistles," *The Journal of the Acoustical Society of America* **94**(5), 2497-2506.

Cambridge University Engineering Department. (**2002**). *Hidden Markov Model Toolkit (HTK) Version 3.2.1 User's Guide*. Cambridge, MA.

Campbell, G. S., Gisiner, R. C., Helweg, D. A., and Milette, L. L. (**2002**). "Acoustic identification of female Steller sea lions (*Eumetopias jubatus*)," *The Journal of the Acoustical Society of America* **111**(6), 2920-2928.

Campbell Jr., J. P. (**1997**). "Speaker recognition: a tutorial," *Proceedings of the IEEE* **85**, 1437-1462.

Charrier, I., Mathevon, N., and Jouventin, P. (**2002**). "How does a fur seal mother recognize the voice of her pup? An experimental study of *Arctocephalus tropicalis*," *The Journal of Experimental Biology* **205**, 603-612.

Charrier, I., Mathevon, N., and Jouventin, P. (**2003**). "Fur seal mothers memorize subsequent versions of developing pups' calls: Adaptation to long-term recognition or evolutionary by-product?," *Biological Journal of the Linnean Society* **80**, 305-312.

Chengalvarayan, R., and Deng, L. (**1997**). "Use of generalized dynamic feature parameters for speech recognition," *IEEE Transactions on Speech and Audio Processing* **5**(3), 232-242.

Chesmore, E. D. (**2001**). "Application of time domain signal coding and artificial neural networks to passive acoustical identificaiton of animals," *Applied Acoustics* **62**, 1359-1374.

Cleveland, J., and Snowdon, C. T. (**1982**). "The complex vocal repertoire of the adult cotton-top tamarin (*Saguinus oedipus oedipus*)," *Z. Tierpsychol.* **58**, 231-270.

Darden, S., Dabelsteen, T., and Pedersen, S. B. (**2003**). "A potential tool for swift fox (Vulpes velox) conservation: Individuality of long-range barking sequences," *Journal of Mammalogy* **84**(4), 1417-1427.

Davis, S. B., and Mermelstein, P. (**1980**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **28**(4), 357-366.

Dayton, L. (**1990**). "Killer whales communicate in distinct 'dialects'," *New Science* **125**, 35.

Deller, J. R., Proakis, J. G., and Hansen, J. H. L. (**1993**). *Discrete-Time Processing of Speech Signals*. New York, NY: Macmillian Publishing Company.

Dudley, H. W. (**1940**). "The carrier nature of speech," *Bell Systems Technical Journal* **19**, 495-513.

Durbin, L. S. (**1998**). "Individuality in the whistle call of the asiatic wild dog *Coun alpinus*," *Bioacoustics* **9**, 197-206.

Egnor, S. E. R., Dingle, K., Guvench, A., Hicks, C., and Hauser, M. D. (**2003**). "The Lombard effect in the cotton-top tamarin (*Saguinus Oedipus*)," *Proceedings of First International Conference on Acoustic Communication by Animals*, University of Maryland, College Park, MD, July 27-30, 2003.

Faucher, A. (**1988**). *The vocal repertoire of the St. Lawrence Estuary population of Beluga whale (Delphinapterus leucas) and its behavioral, social and environmental contexts.* Masters Thesis, Dalhousie University, Halifax, NS, Canada.

Fitch, W. T. (**2003**). "Mammalian vocal production: Themes and variation," *Proceedings of First International Conference on Acoustic Communication by Animals*, University of Maryland, College Park, MD, July 27-30, 2003. 81-82.

Flanagan, J. L. (**1965**). *Speech Analysis and Perception, 2nd edition*. Springer-Verlag.

Fletcher, H. (**1940**). "Auditory patterns," *Reviews of Modern Physics* **12**, 47-65.

Forney, G. D. (**1973**). "The Viterbi Algorithm," *Proceedings of the IEEE* **61**, 268-278.

Fristrup, K. M., and Watkins, W. A. (**1992**). *Characterizing Acoustic Features of Marine Animal Sounds* (Technical Report WHOI-92-04). Woods Hole, MA: Woods Hole Oceanographic Institution.

Fristrup, K. M., and Watkins, W. A. (**1994**). *Marine Animal Sound Classification* (Technical Report WHOI-94-13). Woods Hole, MA: Woods Hole Oceanographic Institution.

Ghosh, J., Deuser, L. M., and Beck, S. D. (**1992**). "A neural network based hybrid system for detection, characterization, and classification of short-duration oceanic signals," *IEEE Journal of Ocean Engineering* **17**(4), 351-363.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," *Hearing Research* **47**, 103-138.

Goldman, J. A., Phillips, D. P., and Fentress, J. C. (**1995**). "An acoustic basis for maternal recognition in timber wolves (*Canis lupus*)?," *The Journal of the Acoustical Society of America* **97**(3), 1970-1973.

Graham, L. H., Bolling, J., Miller, G., Pratt-Hawkes, N., and Joseph, S. (**2002**). "An enzyme-immunoassay for the determination of luteinizing hormone in the serum of African elephants," *Zoo Biology* **21**, 403-408.

Graham, L. H., Schwarzenberger, F., Mostl, E., Galama, W., and Savage, A. (**2001**). "A versatile enzyme-immunoassay for the determination of progestagens in feces and serum," *Zoo Biology* **21**, 227-236.

Greenwood, D. D. (**1961**). "Critical bandwidth and the frequency coordinates of the basilar membrane," *The Journal of the Acoustical Society of America* **33**(10), 1344-1356.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species--29 years later," *The Journal of the Acoustical Society of America* **87**(6), 2592-2605.

Hagan, M. T., Demuth, H. B., and Beale, M. (**1996**). *Neural Network Design*. Boston, MA: PWS Publishing Company.

Harper, M. P., Johnson, M. T., Jamieson, L. H., Hockema, S. A., and White, C. M. (**1999**). "Interfacing a CDG parser with an HMM word recognizer using word graphs," *Proceedings of 1999 International Conference on Acoustics, Speech, and Signal Processing*. 733-736.

Haykin, S. S. (**2002**). *Adaptive Filter Theory, 4th Edition*. Upper Saddle River, NJ: Prentice-Hall.

Heffner, R. S., and Heffner, H. E. (**1982**). "Hearing in the elephant (*Elephas maximus*): Absolute sensitivity, frequency discrimination, and sound localization," *Journal of Comparative and Physiological Psychology* **96**(6), 926-944.

Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis for speech recognition," *The Journal of the Acoustical Society of America* **87**(4), 1738-1752.

Hunt, M. J. (**1999**). "Spectral signal processing for ASR," *Proceedings of 1999 International Workshop on Automatic Speech Recognition and Understanding*.

Insley, S. J. (**1992**). "Mother-offspring separation and acoustic stereotypy: A comparison of call morphology in two species of pinnipeds," *Behaviour* **120**(1-2), 103-122.

Insley, S. J. (**2000**). "Long-term vocal recognition in the northern fur seal," *Nature* **406**, 404-405.

Insley, S. J. (**2001**). "Mother-offspring vocal recognition in northern fur seals is mutual but asymmetrical," *Animal Behaviour* **61**, 129-137.

Insley, S. J., Paredes, R., and Jones, I. L. (**2003**). "Sex differences in razorbill *Alca torda* parent-offspring vocal recognition," *The Journal of Experimental Biology* **206**, 25-31.

Itakura, F. (**1975**). "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**(1), 67-72.

Johnson, C. S., McManus, M. W., and Skaar, D. (**1989**). "Masked tonal hearing thresholds in the beluga whale," *The Journal of the Acoustical Society of America* **85**(6), 2651-2654.

Johnson, M. T., Harper, M. P., and Jamieson, L. H. (**1998**). "Interfacing acoustic models with natural language processing systems," *Proceedings of 1998 International Conference on Speech and Language Processing*. 2419-2422.

Juang, B. H. (**1984**). "On the hidden markov model and dynamic time warping for speech recognition - a unified view," *AT&T Bell Laboratories Technical Journal* **63**(7), 1213-1243.

Ketten, D. R. (**1998**). "A summary of audiometric and anatomical data and its implications fro underwater acoustic impacts," *NOAA Technical Memorandum*.

Kingsley, M. C. S. (**1998**). "Population index estimates for the St. Lawrence Beluga 1973-1995," *Marine Mammal Science* **14**(3), 508-530.

Kogan, J. A., and Margoliash, D. (**1997**). "Automated bird songs recognition using dynamic time warping and hidden Markov models," *The Journal of the Acoustical Society of America* **102**, 3176.

Langbauer Jr., W. R., Payne, K. B., Charif, R. A., Rapaport, L., and Osborn, F. (**1991**). "African elephants respond to distant playbacks of low-frequency conspecific calls," *Journal of Experimental Biology* **157**, 35-46.

Langbauer Jr., W. R., Payne, K. B., Charif, R. A., and Thomas, E. M. (**1989**). "Responses of captive African elephants to playbacks of low-frequency calls," *Canadian Journal of Zoology* **67**, 2604-2607.

Lee, K.-F., and Hon, H.-W. (**1989**). "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing* **37**(11), 1641-1648.

Leong, K. M., Ortolani, A., Burks, K. D., Mellen, J. D., and Savage, A. (**2002**). "Quantifying acoustic and temporal characteristics of vocalizations of a group of captive African elephants (*Loxodonta africana*)," *Bioacoustics* **13**(3), 213-231.

Leong, K. M., Ortolani, A., Graham, L. H., and Savage, A. (**2003**). "The use of low-frequency vocalizations in African elephant (*Locxodonta africana*) reproductive strategies," *Hormones and Behaviour* **43**, 433-443.

LePage, E. L. (**2003**). "The mammalian cochlear map is optimally warped," *The Journal of the Acoustical Society of America* **114**(2), 896-906.

Lombard, E. (**1911**). "L'indication de l'elevation de la voix," *Annales Des Malades de l''creille* **37**(2).

Makhoul, J. (**1975**). "Spectral linear prediction: properties and application," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**, 283-296.

Makhoul, J., and Cosell, L. (**1976**). "LPCW: An LPC vocoder with linear predictive spectral warping," *Proceedings of 1976 International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia. 466-469.

Manabe, K., Sadr, E. I., and Dooling, R. J. (**1998**). "Control of vocal intensity in Budgerigars (*Melopsittacus undulates*): Differential reinforcement of vocal intensity and the Lombard effect," *The Journal of the Acoustical Society of America* **103**, 1190-1198.

Martineau, D. S., DeGuise, S. D., Fournier, M., Shugart, L., Girard, C., Lagace, A., and Beland, P. (**1994**). "Pathology and toxicology of beluga whales from the Saint Lawrence Estuary, Quebec, Canada: Past, present and future," *Science of the Total Environment* **154**, 201-215.

McComb, K., Moss, C., Sayialel, S., and Baker, L. (**2000**). "Unusually extensive networks of vocal recognition in African elephants," *Animal Behaviour* **59**, 1103-1109.

Mellinger, D. K. (**2002**). ISHMAEL: Integrated System for Holistic Multi-channel Acoustic Exploration and Localization. Available at http://cetus.pmel.noaa.gov/cgi-bin/MobySoft.pl.

Mellinger, D. K., and Clark, C. W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation," *The Journal of the Acoustical Society of America* **107**(6), 3518-3529.

Mellinger, D. K., Stafford, K. M., Moore, S. E., Munger, L., and Fox, C. G. (**2004**). "Detection of North Pacific right whale (*Eubalaena japonica*) calls in the Gulf of Alaska," *Marine Mammal Science* **20**(4), 872-879.

Milner, B. (**2002**). "A comparison of front-end configurations for robust speech recognition," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 13-17. Vol. 1, 797-800.

Moon, T. K. (**1996**). "The Expectation-Maximization Algorithm," *IEEE Signal Processing Magazine* **13**(6), 47-60.

Murray, S. O., Mercado, E., and Roitblat, H. L. (**1998**). "The neural network classification of false killer whale (*Pseudorca crassidens*) vocalizations," *The Journal of the Acoustical Society of America* **104**(6), 3626-3633.

Niezrecki, C., Phillips, R., Meyer, M., and Beusse, D. O. (**2003**). "Acoustic detection of manatee vocalizations," *The Journal of the Acoustical Society of America* **114**(3), 1640-1647.

Nonaka, S., Takahashi, R., Enomoto, K., Katada, A., and Unno, T. (**1997**). "Lombard reflex during PAG-induced vocalization in decerebrate cats," *Neuroscience Research* **29**, 283-289.

Oppenheim, A. V., and Schafer, R. W. (**1999**). *Discrete-Time Signal Processing.* Upper Saddle River, NJ: Prentice-Hall.

Owren, M. J., Seyfarth, R. M., and Cheney, D. L. (**1997**). "The acoustic features of vowel-like grunt calls in chacma baboons (*Papio cyncephalus ursinus*): implications for production processes and functions," *The Journal of the Acoustical Society of America* **101**(5), 2951-2963.

Padmanabhan, M., and Picheny, M. (**2002**). "Large-vocabulary speech recognition algorithms," *IEEE Computer* **35**(3), 42-50.

Patterson, R. D. (**1976**). "Auditory filter shapes derived with noise stimuli," *The Journal of the Acoustical Society of America* **59**, 640-659.

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., and Milroy, R. (**1982**). "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold," *The Journal of the Acoustical Society of America* **72**, 1788-1803.

Picone, J. (**1990**). "Continuous speech recognition using hidden Markov models," *IEEE Acoustics, Speech, and Signal Processing Magazine* **7**(3), 26-41.

Poole, J. H., Payne, K. B., Langbauer Jr., W. R., and Moss, C. J. (**1988**). "The social contexts of some very low frequency calls of African elephants," *Behavioral Ecology and Sociobiology* **22**, 385-392.

Potash, L. M. (**1972**). "A signal detection problem and a possible solution in Japanese quail," *Animal Behaviour* **20**, 192-195.

Potter, J. R., Mellinger, D. K., and Clark, C. W. (**1994**). "Marine mammal call discrimination using artificial neural networks," *The Journal of the Acoustical society of America* **96**(3), 1255-1262.

Rabiner, L. R. (**1989**). "Tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE* **77**, 257-286.

Rabiner, L. R., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall.

Recchia, C. A. (**1994**). *Social Behavior of Captive Belugas, Delphinapterus leucas*. Ph.D. Thesis (WHOI Report WHOI-94-03), Massachusetts Institute of Technology / Woods Hole Oceanographic Institute.

Reynolds, D. A. (**2002**). "An overview of automatic speaker recognition technology," *Proceedings of 2002 International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL. Vol. 4, 4072-4075.

Riede, T., and Zuberbühler, K. (**2003**). "The relationship between acoustic structure and semantic information in Diana monkey alarm vocalizations," *The Journal of the Acoustical Society of America* **114**(2), 1132-1142.

Robinson, D. W., and Dadson, R. S. (**1956**). "A redetermination of the equal-loudness relations for pure tones," *British Journal of Applied Physics* **7**, 166-181.

Roe, D. B., and Wilpon, J. G. (**1993**). "Whither speech recognition: the next 25 years," *IEEE Communications Magazine* **31**(11), 54-62.

Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**, 43-49.

Santivo, S., and Galimberti, F. (**2000**). "Bioacoustics of southern elephant seals. II. Individual and geographical variation in male aggressive vocalisations," *Bioacoustics* **10**, 287-307.

Scheifele, P. M. (**2003**). *Investigation into the response of the auditory and acoustic communication systems in the beluga whale (Delphinapterus leucas) of the St. Lawrence River estuary to noise, using vocal classification*. Ph. D. Dissertation, University of Connecticut, Hartford, CT.

Scheifele, P. M. (**2004**). Personal communication.

Schön, P.-C., Puppe, B., and Manteuffel, G. (**2001**). "Linear prediction coding analysis and self-organizing feature map as tools to classify stress calls of domestic pigs (*Sus scrofa*)," *The Journal of the Acoustical Society of America* **110**(3), 1425-1431.

Schroeder, M. R. (**1977**). *Recognition of Complex Acoustic Signals, Life Sciences Research Report 5*. Edited by T. H. Bullock. Berlin: Abakon Verlag.

Silverman, H. F., and Morgan, D. P. (**1990**). "The application of dynamic programming to connected speech recognition," *IEEE Acoustics, Speech, and Signal Processing Magazine* **7**, 6-25.

Sinott, J. M., Stebbins, W. C., and Moody, D. B. (**1975**). "Regulation of voice amplitude by the monkey," *The Journal of the Acoustical Society of America* **58**(2), 412-414.

Sjare, B. L., and Smith, T. G. (**1986a**). "The relationship between behavioral activity and underwater vocalizations of the white whale, *Delphinapterus leucas*.," *Canadian Journal of Zoology* **64**, 2824-2831.

Sjare, B. L., and Smith, T. G. (**1986b**). "The vocal repertoire of white whales, *Delphinapterus leucas*, summering the Cunningham Inlet, Northwest Territories," *Canadian Journal of Zoology* **64**, 407-415.

Soltis, J., Leong, K. M., and Savage, A. (**2005**). "African elephant vocal communication II: How many rumbles are there?," *in preperation*.

Sousa-Lima, R. S., Paglia, A. P., and Da Fonseca, G. A. B. (**2002**). "Signature information and individual recognition in the isolation calls of Amazonian manatees, *Trichechus inunguis*," *Animal Behaviour* **63**, 301-310.

Stevens, S. S. (**1957**). "On the psychophysical law," *Psychological Review* **64**, 153-181.

Stevens, S. S., and Volkmann, J. (**1940**). "The relation of pitch to frequency: A revised scale," *American Journal of Psychology* **53**, 329-353.

Stoica, P., and Moses, R. L. (**1997**). *Introduction to Spectral Analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Titze, I. R. (**1994**). *Principles of Voice Communication*. Englewood Cliffs, NJ: Prentice-Hall.

Watkins, W. A., Daher, M. A., DiMarzio, N. A., and Reppucii, G. (**1998**). *Distinctions in sound patterns of calls by killer whales (Orcinus Orca) from analysis of computed sound features* (Technical Report WHOI-98-05). Woods Hole, MA: Woods Hold Oceanographic Institution.

Weisburn, B. A., Mitchell, S. G., Clark, C. W., and Parks, T. W. (**1993**). "Isolating biological acoustic transient signals," *Proceedings of 1993 International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1, 269-272.

Zwicker, E., Flottorp, G., and Stevens, S. S. (**1957**). "Critical band width in loudness summation," *The Journal of the Acoustical Society of America* **29**(5), 548-557.

Zwicker, E., and Terhardt, E. (**1980**). "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *The Journal of the Acoustical Society of America* **68**(5), 1523-1525.

APPENDIX A – DERIVATION OF EQUATIONS FROM ERB DATA

The constant values A, a, and k for the Greenwood warping function can be derived from an ERB function of the form

$$ERB = \alpha(\beta f + \delta) \tag{A.1}$$

by taking the integral of the inverse as follows (Zwicker and Terhardt, 1980):

$$f_p = \int \frac{1}{\alpha(\beta f + \delta)}, \tag{A.2}$$

$$f_p = \frac{1}{\alpha}\int \frac{1}{\beta f + \delta},$$

$$f_p = \frac{1}{\alpha\beta}\ln(\beta f + \delta) + C. \tag{A.3}$$

It is usually desirable that f_p=0 when f=0. For this to be the case, the integration constant C must be set to 0 and δ must be 1. The base of the logarithm is then changed to 10 to match the base in equation 4.6,

$$f_p = \frac{1}{\alpha\beta\log_{10}(e)}\ln(\beta f + \delta). \tag{A.4}$$

This equation is in the same form as equation 4.6, and the constant values can be read directly as

$$A = \frac{1}{\beta}, \tag{A.5}$$

$$a = \alpha\beta\log_{10}(e), \text{ and} \tag{A.6}$$

$$k = \delta. \tag{A.7}$$

APPENDIX B – DERIVATION OF EQUATIONS FROM APPROXIMATE HEARING RANGE

The constant values A, a, and k for the Greenwood warping function can be derived using the assumption from LePage (2003) that the optimal value of k=0.88 and estimates of the hearing range of the species, f_{min} to f_{max}. The constant A can be solved for using the constraint $F_p(f_{min})$=0:

$$0 = (1/a)\log_{10}(f_{min}/A + k), \tag{B.1}$$

$$0 = \log_{10}(f_{min}/A + k),$$

$$1 = f_{min}/A + k,$$

$$1 - k = f_{min}/A, \text{ and}$$

$$A = \frac{f_{min}}{1-k}. \tag{B.2}$$

The constant a can then be determined using the constraint $F_p(f_{max})$=1 and equation B.2,

$$1 = (1/a)\log_{10}(f_{min}/A + k) \tag{B.3}$$

and then solving for a,

$$a = \log_{10}(f_{min}/A + k). \tag{B.4}$$

APPENDIX C – DERIVATION OF MAXIMUM NUMBER OF FILTERS

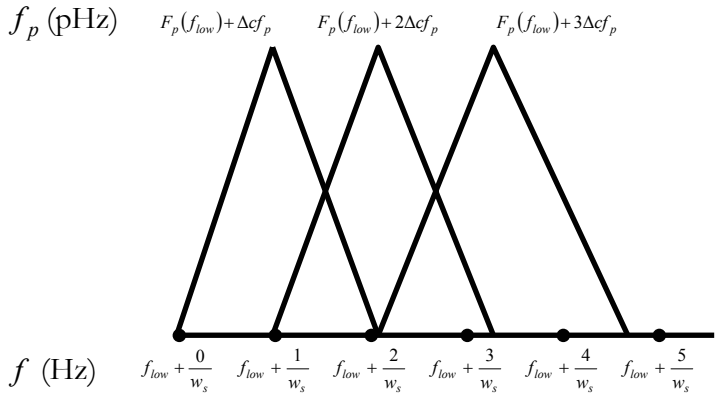


Figure C.1 – Diagram of Filter Bank

Using Figure C.1, it can be seen that for at least 1 point to contribute significantly to the first filter's energy, the value of the second filter's center frequency must be greater than the frequency represented by the second point in the spectrum. Generalizing this constraint, the second filter's center frequency must be greater than the frequency represented by the n+1th point in the spectrum for n points to contribute significantly to the first filter's energy. This can be represented by:

$$F_p^{-1}\left(F_p\left(f_{low}\right)+2\Delta cf_p\right) > f_{low}+\gamma/w_s\ , \tag{C.1}$$

where γ is n+1. Substituting the equation for Δcf_p yields

$$F_p^{-1}\left(F_p\left(f_{low}\right)+2\left[\frac{F_p\left(f_{high}\right)-F_p\left(f_{low}\right)}{n_f+1}\right]\right) > f_{low}+\gamma/w_s\ . \tag{C.2}$$

Then, taking the F_p of both sides,

$$F_p\left(f_{low}\right)+2\left[\frac{F_p\left(f_{high}\right)-F_p\left(f_{low}\right)}{n_f+1}\right] > F_p\left(f_{low}+\gamma/w_s\right), \tag{C.3}$$

and solving for n_f yields

$$2\left[\frac{F_p\left(f_{high}\right)-F_p\left(f_{low}\right)}{n_f+1}\right] > F_p\left(f_{low}+\gamma/w_s\right)-F_p\left(f_{low}\right),$$

$$\frac{1}{n_f+1} > \frac{F_p\left(f_{low}+\gamma/w_s\right)-F_p\left(f_{low}\right)}{2\left(F_p\left(f_{high}\right)-F_p\left(f_{low}\right)\right)}, \text{ and finally}$$

$$n_f < \frac{2\left(F_p\left(f_{high}\right)-F_p\left(f_{low}\right)\right)}{F_p\left(f_{low}+\gamma/w_s\right)-F_p\left(f_{low}\right)}-1. \qquad \text{C.4}$$

Marquette University

This is to certify that we have examined
this copy of the
dissertation by

Patrick J. Clemins, B.S., M.S.

and have found that it is complete
and satisfactory in all respects.

This dissertation has been approved by:

_____
Michael T. Johnson, Ph.D., P.E.
Dissertation Director, Department of Electrical and Computer Engineering

_____
George F. Corliss, Ph.D.
Committee Member

_____
James A. Heinen, Ph.D.
Committee Member

_____
Richard J. Povinelli, Ph.D., P.E.
Committee Member

_____
Anne Savage, Ph.D.
Committee Member

Approved on

_____