Bat Species Identification from Zero Crossing and Full Spectrum Echolocation Calls using HMMs, Fisher Scores, Unsupervised Clustering and Balanced Winnow Pairwise Classifiers

Ian Agranat, Wildlife Acoustics, Inc. Concord, Massachusetts, U.S.A.

Abstract

A new classification technique for the identification of bats to species from their echolocation calls is presented. Three different datasets are compiled and split in half for training and testing classifiers. Combined, the data include 9,014 files (bat passes) with 226,432 candidate calls (pulses or extraneous noise) representing 22 different species of bats found in North America and the United Kingdom. Some files are of high quality consisting of hand-selected search phase calls of tagged free flying bats while others are from a variety of field conditions including both active (attended) and passive (unattended) recordings made with a variety of zero crossing and full spectrum recording equipment from multiple vendors.   Average correct classification rates for the three datasets on test data are 100.0%, 97.9%, and 88.8% respectively with an average of 92.5%, 72.2%, and 39.9% of all files identified to species. Most importantly, classifiers in the third dataset for two species of U.S. endangered bats, *Myotis sodalis* (MYSO) and *Myotis grisescens* (MYGR) have a correct classification rate of 100% and 98.6% respectively and identify 67.4% and 93.8% of all files to species suggesting that the classifiers are well suited to the accurate detection of these endangered bats.

Introduction

Populations of bats are commonly monitored acoustically because many species echolocate while foraging at night. Echolocation calls are typically recorded using zero crossing detectors or full spectrum recorders at high sample rates, and the resulting recordings are then analyzed with software to display the frequency modulated sweeps common to many bat species. Identifying bat species from their echolocation calls is desirable for management of biodiversity and compliance with environmental regulations. Human experts have proven that it is possible in many cases to identify bats by analysis of their echolocation calls. However, variation at several levels makes some species indistinguishable (Barclay, 1999). Many bats produce a wide range of calls to adapt to their physical surroundings, including the presence of other bats, and these calls often converge across species to very similar call types making identification difficult. Search phase calls are best suited for the acoustical identification of bats because they are the most commonly encountered in the field and have been shown to have species-specific characteristics (Allen, Burt, and Miller, 2007). However, field recordings collected from unattended passive monitoring sites will likely have a wide variety of call types present (e.g. clutter, feeding buzz, etc.) and quality (e.g. near and far, insect noise, and echoes).

There have been several efforts to develop algorithms for the automatic classification of bat calls including (Parsons and Jones, 2000), (Fukui, Agetsuma, and Hill, 2004), (Skowronski and Harris, 2006), (Corcoran, 2007), (Redgwell, Szewczak, Jones, and Parsons, 2009), (Britzke, E.R., J. Duchamp, R.S. Swhiart, K.M. Murray, and L.W. Robbins, 2011) et. al. Prior published efforts are generally limited in scope attempting to classify from among only a dozen or so individual species using a relatively small number of hand-labeled search phase calls. Their performance when presented with the likely variety of call types and quality typical of unattended field recordings is unknown. Prior methods generally involve the extraction of discrete parameters (e.g. mean frequency, etc.) from each call pulse to form a feature vector, and then use any number of well-known techniques to classify to species including Discriminant Function Analysis, Artificial Neural Networks, Random Forest, and Support Vector Machines. While these parameters have demonstrated good classification rates on high quality search phase calls, consistent determination of parameters in noisy environments can be challenging and these parameters may not contain sufficient discriminant information to separate a larger variety of call types into classes. Finally each of these methods was developed specifically for analysis of either zero crossing or full spectrum recordings, not both, limiting the choices of recording systems that can be utilized by prior methods (Allen, C.R., S.E. Romeling, and L.W. Robbins, 2011).

The objective of this research is to develop new techniques for accurately classifying the echolocation calls of bats on a large scale in the presence of a wide variety of call types, field conditions and recording technologies. The author's prior research in animal vocalization classification (Agranat, 2009) provides an imperfect starting point and motivation to find improvements with the hope that these new techniques may also find applications in other acoustic classification problems including birds, frogs, and cetaceans.

Full Spectrum vs. Zero Crossing

Full Spectrum ultrasonic recordings are digital recordings made at high sample rates, typically 200-500kHz, to record bat calls up to 100-150kHz. These recordings are analyzed by Fourier transforms to generate spectrograms representing the frequency sweep of echolocation calls including harmonic details and the power distribution of the signal. Zero crossing recordings operate in the time domain and count the delay between successive zero-crossings of the signal above some noise threshold. The time between zero-crossings, or more commonly between a fixed number of zero crossings known as the division ratio, is recorded. Zero crossing analysis can derive the frequency sweep of the echolocation call through time representing the strongest frequency components of the call. No amplitude or harmonic structure is present.

In theory, the full spectrum recordings should be better suited to the task of classification because more data are available. However, this is not necessarily true. The added dimension of amplitude information can be very sensitive to the frequency response of the ultrasonic transducer, the effect of weatherproofing, the distance of the bat, interference from echoes, and other noisy factors that may in fact interfere with reliable classification. While it may be true that amplitude information could be a critical component to accurately classifying some species, human experts have been able to reliably identify many species from zero crossing recordings without the benefit of this extra information. Additionally, the prior efforts report similarly high classification rates when either technique was used. Since the objective is to work with as much data as possible, it is desirable to develop techniques that work with both full spectrum and zero crossing technologies, and zero crossing is the least common denominator. A full spectrum recording can be converted to zero crossing, but not the other way around. So, this solution is based on zero crossings of the echolocation call, but can work with full spectrum recordings from which the zero crossing information is extracted.

Full spectrum recordings enjoy advantages over native zero crossing recordings in that zero crossing information can be extracted from full spectrum recordings that would not be possible directly from a native zero crossing recorder. Consider the case of a weak bat signal or a bat signal in the presence of insect noise. A native zero-crossing detector may not be capable of detecting a weak signal against broadband background noise. On the other hand, a full spectrum recording can be manipulated in the frequency domain by applying noise reduction, echo cancellation and band-pass filters to detect, extract and enhance the narrowband signal representing the echolocation calls of bats. These techniques are beyond the scope of this paper but are embodied in Kaleidoscope^{TM} software from Wildlife Acoustics.

Incomplete Knowledge and Imperfect Data

The objective of this research is to build a large scale classifier for the accurate identification of many species (worldwide in scope) with confidence by running trials against a massive library of both full spectrum and zero-crossing recordings identified to species from numerous sources. A subset of these recordings would be used for training classifiers and the remaining recordings would be used to verify their performance. This leads immediately to some challenges. First, can the accuracy of recording file labels be trusted? And second, how are the individual echolocation calls that are most suitable for classification (e.g. search phase calls) in each file determined?

To answer the first question, it is unlikely that the data labels will be error free. Certainly some contributions to the collection may include files that were misclassified either from confusion about the identity of a call or clerical error in organizing the data. It is also possible that some files may contain calls from more than one species. Overall, file labels will generally be accurate, but methods should be robust against some portion of the labels being inaccurate.

To answer the second question, remember that prior efforts generally used hand-labeled search-phase calls. Not only would hand labeling each individual call be impractical on such a large scale, but this in fact may be undesirable if the underlying assumption that search-phase calls are the most important for classification turns out not to be true. Rather than make assumptions about the structure of bat calls and their importance in classification, they should be automatically discovered through machine learning.

The algorithms described below build classifiers from a large library consisting of both full spectrum and zero crossing recording files with good but imperfect file-level species labels automatically with no human intervention.

Parameterization of the Echolocation calls of Bats

Prior work has generally relied on a series of discrete parameters to describe the frequency modulated sweep of a bat's echolocation call. This approach is limiting in that call parameters cannot always be extracted consistently from noisy environments, and may miss important subtleties present in the call needed for discrimination. Some pairs of species such as *Myotis lucifugus* and *Myotis sodalis* have significant overlap in discrete parameters such as call duration, characteristic frequency, start slope, slope at characteristic frequency, and cumulative normalized slope such that these species cannot be differentiated (Szewczak, 2011).

The following figures illustrate this point with scatter plots of characteristic slope (Sc) vs. call duration, initial slope (S1) vs. call duration and call characteristic frequency (Fc) vs. call duration from *Myotis lucifugus* (MYLU) and *Myotis sodalis* (MYSO) noting that the former is relatively common in the United States while the latter is endangered. Notice that there is significant overlap between these two species and they cannot be separated on the basis of these parameters alone. The MYLU call distribution includes longer duration calls than MYSO and thus longer MYLU calls can be identified. However, shorter duration MYLU calls are nearly indistinguishable from MYSO calls.

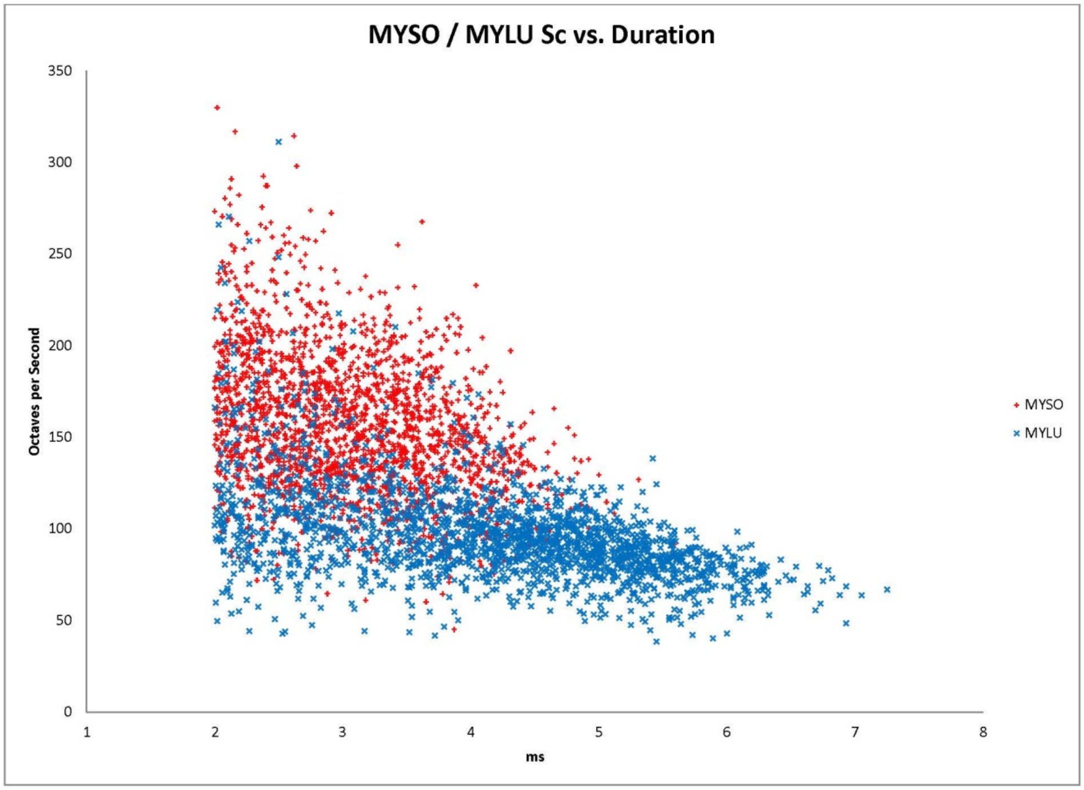

Figure 1: MYSO/MYLU Discrimination Characteristic Slope vs. Duration

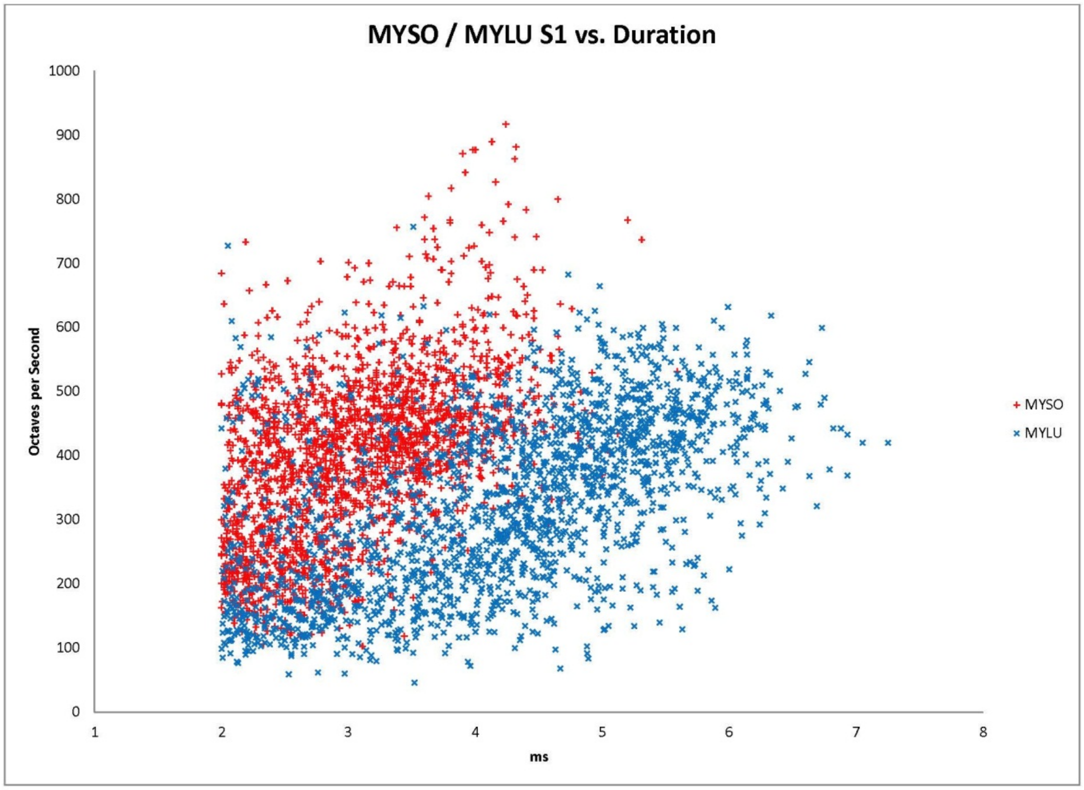

Figure 2: MYSO/MYLU Discrimination Initial Slope vs. Duration

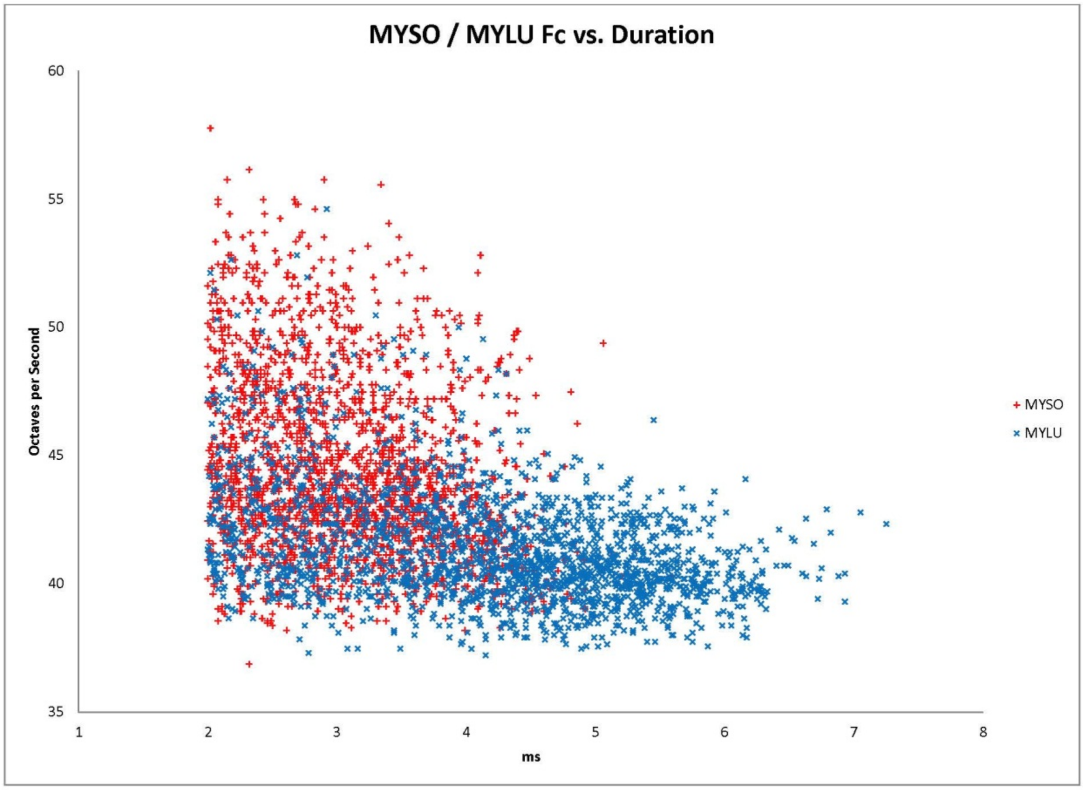

Figure 3: MYLU/MYSO Discrimination Characteristic Frequency vs. Duration

In many other disciplines including speech recognition, handwriting recognition, face recognition, electrocardiogram analysis, DNA sequencing, etc., the Hidden Markov Model (HMM) has proven to be a powerful generative model for representing variable length observation sequences. HMMs are the basis of the author's prior work on bird song classification as well. In simple terms, an HMM is a generative model $\boldsymbol{\varphi}$ that has a probability of generating a sequence of observations. The model is represented as a collection of states $\boldsymbol{X_i}$ with each state having a probability $\boldsymbol{b_{ij}}$ of emitting an observation $\boldsymbol{y_j}$, and transition probabilities $\boldsymbol{a_{ij}}$ of moving from one state $\boldsymbol{X_i}$to the next state $\boldsymbol{X_j}$ at each step $\boldsymbol{t}$. The observations can model discrete symbols or continuous functions, the latter commonly represented as a $\boldsymbol{d}$-dimensional Gaussian Mixture Model (GMM) with mixing coefficients $\boldsymbol{m_{ik}}$, mean vectors μ$_{\boldsymbol{ik}}$ and covariance matrixes $\boldsymbol{\Sigma_{ik}}$.

In the GMM case, the emission probability is given by the Gaussian probability density function:

$$b_{ik}(y) = m_{ik}\ argmax_k[\ (2\pi)^{-\frac{d}{2}}|\Sigma_{ik}|^{-\frac{1}{2}}e^{-\frac{1}{2}(y-\mu_{ik})'\Sigma_{ik}{}^{-1}(y-u_{ik})}\ ]$$

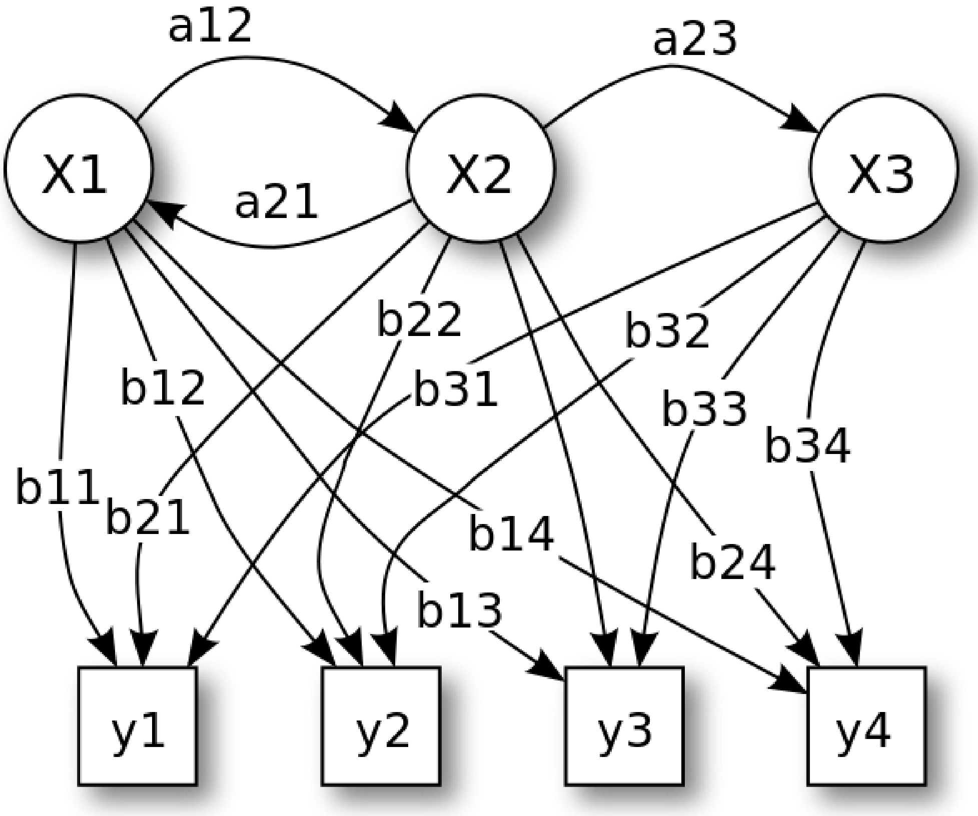

Figure 4: Hidden Markov Model (image courtesy of Wikipedia)

Given an HMM $\boldsymbol{\varphi}$ and an observation sequence $\boldsymbol{y}$, the prior probability $\boldsymbol{P(y|\varphi)}$ can be calculated by using the iterative Viterbi algorithm:

$$V_{0,k} = b(y_0|k)\ \pi_k$$

$$V_{t,k} = b(y_t|k)\ argmax_{x\in X}\ (a_{x,k}V_{t-1,x})$$

$$P(y|\varphi) = argmax_{x\in X}(V_{T,x})$$

In zero-crossing recordings, the frequency modulated sweep of the bat’s echolocation call through time is represented as a series of points, or “dots”, one dot for every N zero crossings in the signal. The time between dots is stored in the recording. For analysis, the frequency and time position of each dot can be easily calculated, and this is how a plot of a bat’s frequency sweep through time can be generated. It is natural to think of the sequence of dots from a zero-crossing recording as an observation sequence, with each dot corresponding to an observation.

The approach described herein uses HMMs to model the variable length sequence of dots normalized to the same division ratio to represent the echolocation calls of bats. To help with model alignment and because the slope of the frequency sweep is known to be one of the important parameters used in prior efforts, a 2-dimensional feature vector modeled as a GMM for each state is used. One dimension represents the frequency of each dot, and the second dimension represents the change of frequency from the previous dot. These values are normalized to unit variance across the training data, and a diagonal covariance matrix is used for convenience.

A Hidden Markov Model can then be used to model bat echolocation calls. A reasonable estimate for an initial model would be one state for each dot with states having left-to-right transitions probabilities through the duration of the call. In practice, clustering observations into a smaller number of states using K-means also works well and is computationally more efficient.

By using HMMs to model echolocation calls, every detail of the call's frequency sweep through time is captured without relying on discrete parameters to approximate isolated portions of the call. Hidden Markov Models have also been proven in other applications to be robust against noise and variation.

Discovering Call Types with Unsupervised Clustering

Given a large collection of files, each containing several echolocation calls and labeled with some imperfect degree of accuracy as belonging to a particular species, one objective is to automate the process of discovering those call types that are common to the species and therefore good candidates on which to base the classifiers. Another objective is to be robust against the inclusion of some files that are incorrectly labeled and noise incorrectly extracted as calls.

Various techniques for clustering HMMs are described by (Butler, 2003), (Chen, Man, and Nefian, 2005), et. al. used for analysis of speech, whale song, face recognition and other applications and are ideal for clustering variable-length time-series data. Training files are selected with the same species label and individual echolocation calls are extracted (or possibly other noise fragments erroneously extracted from field recordings) on which to perform agglomerative clustering. In one approach, each of these echolocation calls is represented by a single HMM. All of the HMMs are compared pairwise and the two most similar HMMs are merged into a single HMM. This process is repeated until the desired number of clusters has been realized. Unfortunately, this approach does not scale well for large data sets. Instead, a single HMM is trained with all of the calls, and then the inner product of Fisher scores is used to measure the similarity of calls.

The Fisher score is defined as:

$$U_y = \nabla\varphi \; log \; P(y|\varphi)$$

In the context of HMMs, the Fisher score represents how an observation sequence fits the model with respect to each of the model's underlying parameters. The Fisher score is the gradient of the log probability of the observation sequence with respect to each of the model parameters. Two variable length observation sequences can be compared by taking the inner product of their Fisher scores. The angular distance between these two observations in hyperspace is thus:

$$cos^{-1}(Ux \cdot Uy)$$

The algorithms perform agglomerative clustering by grouping the most similar calls together. This generally results in a small number of large clusters per species with a number of smaller clusters and un-clustered outliers. The larger clusters represent the most common call types present throughout a large portion of the training data (e.g. search phase calls). Smaller clusters may represent uncommon or highly variable calls, or even groups of misclassified calls that are similar to each other but don't actually belong in the data set. The largest cluster by membership is included among the call types of interest. Additionally, any other similarly large clusters with at least 50% of the membership of the largest cluster are also included. In this way, a handful of HMMs are formed that are trained on the most common call types of a given species, and outliers are explicitly discarded.

To illustrate the power of this approach, consider the MYLU/MYSO classification problem described previously that could not be easily solved using simple parametric measurements. This new clustering technique applied to the second dataset (Midwest search phase calls) discovered one MYLU cluster (with 332 member calls) and one MYSO cluster (with 456 member calls). HMMs are trained with left-to-right topology with a number of states equal to the average sequence length. A plot of the mean frequency with error bars of each state is shown for these clusters.

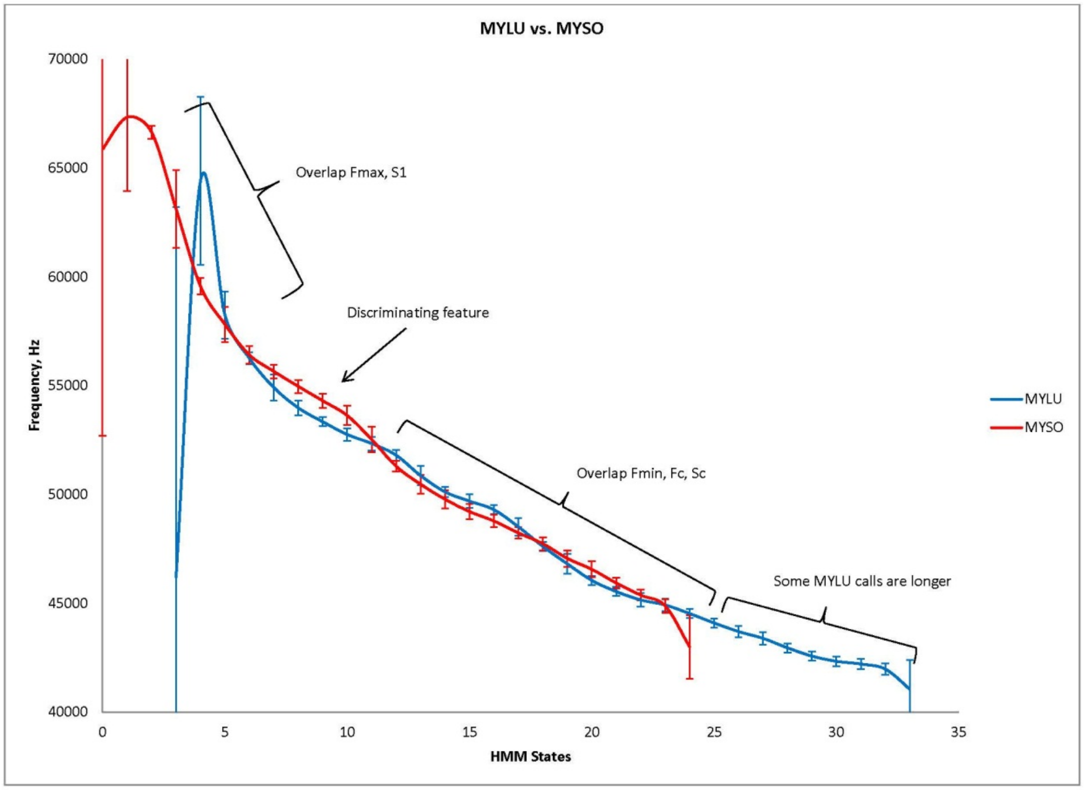

Figure 5: HMMs separating MSYO and MYLU.

Note that the MYLU cluster and the MYSO cluster overlap significantly in several places, but diverge slightly in states 5-10. Traditional parameters like the characteristic frequency (Fc), characteristic slope (Sc), initial slop (S1) and others measure portions of the call where discrimination is not possible. On the other hand, HMMs allow us to model the entire call holistically and can tease apart subtle differences to greatly improve classification performance. The HMMs automatically discover differences in the calls encoded in the model parameters. Analysis of these model parameters suggests that these two species can be differentiated from a subtle difference in the slope of the frequency trajectory at around 55kHz. This discovered feature does not contribute significantly to the traditional discrete parameters used by prior methods and solves the classification problem where others have struggled.

Classification of Calls

Given an unclassified observation sequence representing an unknown bat call and a set of trained HMMs representing different species classes, a simple method of classification is to calculate the prior probability that each HMM generated the sequence and choose the class with the highest probability. This is essentially what the author's prior work on birdsong classification was based on. However, it turns out that this technique does not produce reliable results. Hidden Markov Models are good generative models and are optimized to maximize the prior probabilities of the training sequences, but this is not the same as maximizing the discrimination between classes. Classification can break down in a number of scenarios. In the author's prior work, some models with loop states and broader GMM variances became "greedy" allowing unrelated sequences to produce high scores. Additionally, short duration sequences that fit closely to a portion of a model trained on longer sequences scored high as well. When two classes are similar, the optimization for maximizing the probabilities of the training data without regard to maximizing the margin between classes can result in high rates of misclassification.

Fisher scores can improve discrimination of generative models. Using Fisher scores, a variable length observation sequence applied to a Hidden Markov Model can be transformed into a fixed-length feature vector in high dimensional space with one dimension corresponding to each of the underlying model parameters (e.g. state transition probabilities, means and variances of GMMs, etc.). This technique has been used successfully in many applications including sign language recognition (Aran and Akarun, 2009).

The Fisher score of an observation sequence for an HMM corresponding to the related class reveals which model parameters are most significant in their contribution to the prior probability. The Fisher score of the same sequence for unrelated HMMs can also be revealing in exposing similarities and differences between classes.

The method described uses a Fisher score vector and prior probability for all the included HMM clusters for all species as a classification feature vector given an unknown observation sequence. The number of parameters in each HMM can be quite large. An N-state model would have N initial state probabilities, N^2 state transition probabilities, and N GMMs each containing (in our case) one mixture in two dimensions with a mean and diagonal covariance matrix for a total of $5N+N^2$ parameters. A 16 state model would therefore contribute 336 parameters to the feature vector. Given dozens of species, our feature vector can quickly grow to on the order of 10,000 dimensions. Only a small portion of these dimensions are expected to be significant for discrimination, many dimensions will be noisy, and many dimensions will have no significance (e.g. representing model parameters for seldom visited states or state transitions).

Binary classifiers based on Support Vector Machines are popular tools for handling highly dimensional spaces, but very good results are achieved using the simpler balanced Winnow methods. The Winnow algorithm is capable of rapid convergence on linearly separable data even when few variables are relevant (Kivinen and Warmuth, 1995).

The balanced Winnow algorithm finds a hyper-plane separating two classes in multidimensional space defined by a positive and negative weight vector as follows:

$\sum_{i=0}^{n} (w_i^+ x_i - w_i^- x_i)$ > 0, class c = 1. Otherwise c = -1.

The classifier is trained with an online learning algorithm. When a mistake is made, the weights are updated exponentially as follows:

$$w_{i,t+1}^+ = \frac{w_{i,t}^+ e^{c\beta x_i}}{\sum_{i=0}^{n} w_{i,t+1}^+}$$

$$w_{i,t+1}^- = \frac{w_{i,t}^- e^{-c\beta x_i}}{\sum_{i=0}^{n} w_{i,t+1}^-}$$

After training, the weights of the Winnow hyper-planes can be analyzed to reduce the dimensionality of the problem significantly in the final classifier by eliminating the dimensions that do not contribute significantly to the classification.

Identification of bats to species is a multi-class classification problem solved with binary classifiers capable of separating data sets with linear hyper-planes in highly dimensional space. One approach is a one-against-all strategy in which a binary classifier is trained for each class in an attempt to find a hyper-plane that separates members in the class from members in all other classes. Such an approach if possible has the benefit of excluding noise and other non-identifiable calls from analysis. Unfortunately, it is not likely given the number of classes and the similarity among some classes that a one-against-all approach is solvable with a linear separating hyper-plane.

Instead, an exhaustive pair-wise strategy is used with a binary classifier trained to find a separating hyper-plane between each pair of classes. In addition, one extra class is defined to represent the unidentifiable sequences derived from the discarded HMM clusters. Using this method, Winnow converges on a linear separating hyper-plane with average error less than 3% for all pairwise combinations.

The outputs of the pairwise classifiers are combined using a hinge loss function to predict the winning class with a confidence factor. If the winning class turns out to be the unidentifiable sequence class, then the call is rejected as unknown.

At the file level, there may be a sequence of echolocation calls recorded during a “bat pass” over several seconds. Each call is classified as above with the confidence factors accumulated by species and normalized across the file. The file is considered identified to species if the normalized confidence factors exceed some threshold.

Discussion and Results

Recordings of bats were acquired from four different sources, one in the United Kingdom and three in North America for this research.

The first dataset includes 6 different species from the United Kingdom and contains 232 files (bat passes) from which 3,293 candidate calls (pulses or extraneous noise) were extracted. These were all recorded with full spectrum equipment from Wildlife Acoustics.

The second dataset includes 10 different species of bats found in the Midwestern United States and contains 732 high quality files. These files were recorded in zero crossing with AnaBats from Titley Scientific and were hand selected as containing search phase calls of known bats in free flight, often employing the use of light tags. From these files, 25,159 candidate calls were extracted.

A third dataset combines the second dataset described above with all of the files from the other two sources. Combined, this dataset includes 17 species of bats found throughout North America including 10 species of *Myotis*. The dataset has 8,782 files from which 223,123 candidate calls were extracted. These recordings were made in a variety of field conditions including both active (attended) and passive (unattended) recordings made with a variety of zero crossing and full spectrum recording equipment from multiple vendors. The dataset is not evenly distributed among species with 38.0% of the files representing *Eptesicus fuscus* (EPFU) and only 0.6% representing *Corynorhinus townsendii* (COTO). The table below lists the species present in these three groups.

Zero crossing files were normalized to a division ratio of 8 and full spectrum files were converted to zero crossing with Kaleidoscope. Individual echolocation calls are extracted from each file by looking for chains of between 10-100 dots forming smooth trajectories. After a call is extracted, 50ms of the file is skipped to avoid echoes before looking for the next call. Each call is represented as a sequence of two-dimensional feature vectors, one vector for each dot, with dimensional components representing the frequency and change in frequency normalized to unit variance.

Each dataset is split in half with approximately one half of the files used for training the classifiers and the other half used for testing the classifiers against previously unconsidered recordings.

Table 1: List of Species

Common Name	**Scientific Name**	**Abbreviation**
Western Barbastelle	*Barbastella barbastellus*	BABA
Townsend's Big-Eared Bat	*Corynorhinus townsendii*	COTO
Big Brown Bat	*Eptesicus fuscus*	EPFU
Eastern Red Bat	*Lasiurus borealis*	LABO
Hoary Bat	*Lasiurus cinereus*	LACI
Silver-haired Bat	*Lasionycteris noctivagans*	LANO
California Myotis	*Myotis californicus*	MYCA
Western Small-footed Bat	*Myotis ciliolabrum*	MYCI
Gray Bat	*Myotis grisescens*	MYGR
Keen's/Long eared Myotis	*Myotis keenii/Myotis evotis*	MYKEMYEV
Little Brown Bat	*Myotis lucifugus*	MYLU
Northern long-eared Myotis	*Myotis septentrionalis*	MYSE
Indiana Bat	*Myotis sodalis*	MYSO
Fringed Myotis	*Myotis thysanodes*	MYTH
Long-legged Myotis	*Myotis volans*	MYVO
Yuma Myotis	*Myotis yumanensis*	MYYU
Evening Bat	*Nycticeius humeralis*	NYHU
Common Noctule	*Nyctalus noctula*	NYNO
Tricolored Bat	*Perimyotis subflavus*	PESU
Common Pipistrelle	*Pipistrellus pipistrellus*	PIPI
Soprano Pipistrelle	*Pipistrellus pygmaeus*	PIPY
Greater Horseshoe	*Rhinolophus ferrumequinum*	RHFE
Lesser Horseshoe	*Rhinolophus hipposideros*	RHHI

The classifier is built from the training data as follows: For each species, a Hidden Markov Model is trained on all of the corresponding calls. For each pair of calls corresponding to each species, the inner product of their Fisher scores is used to determine their similarity expressed as an angular separation in hyperspace. Agglomerative clustering is performed by merging calls from the smallest angular separation up to some maximum. This generally results in a few large clusters per species which are kept while discarding the others. The resulting clusters are shown in the tables below:

Table 2: Discovered Clusters (Dataset #1)

Dataset	Cluster	Size	Total Calls	
1	BABA.0	39	73	53%
1	NYNO.0	64	300	21%
1	NYNO.1	89	"	30%
1	PIPI.0	160	840	19%
1	PIPI.1	260	"	31%
1	PIPI.2	231	"	28%
1	PIPY.0	280	379	74%
1	RHFE.0	122	131	93%
1	RHHI.0	75	77	97%
		1,320	1,800	73%

Table 3: Discovered Clusters (Dataset #2)

Dataset	Cluster	Size	Total Calls	
2	EPFU.0	1744	2097	83%
2	LABO.0	840	923	91%
2	LACI.0	262	422	62%
2	LANO.0	396	425	93%
2	MYGR.0	553	1453	38%
2	MYLU.0	1085	1168	93%
2	MYSE.0	1026	1195	86%
2	MYSO.0	501	1577	32%
2	MYSO.1	727	"	46%
2	NYHU.0	546	2251	24%
2	NYHU.1	788	"	35%
2	NYHU.2	516	"	23%
2	PESU.0	660	1504	44%
2	PESU.1	335	"	22%
		9,979	13,015	77%

Table 4: Discovered Clusters (Dataset #3)

Dataset	**Cluster**	**Size**	**Total Calls**	
3	COTO.0	154	451	34%
3	COTO.1	152	"	34%
3	EPFU.0	3,861	38,036	10%
3	EPFU.1	5,913	"	16%
3	EPFU.2	3,539	"	9%
3	EPFU.3	3,896	"	10%
3	EPFU.4	5,583	"	15%
3	EPFU.5	5,358	"	14%
3	LABO.0	2,837	9,334	30%
3	LABO.1	2,361	"	25%
3	LACI.0	968	3,940	25%
3	LACI.1	1,891	"	48%
3	LANO.0	592	1,521	39%
3	LANO.1	489	"	32%
3	MYCA.0	1,098	2,119	52%
3	MYCI.0	1,012	1,906	53%
3	MYCI.1	666	"	35%
3	MYGR.0	1,366	1,453	94%
3	MYKEMYEV.0	1,777	7,335	24%
3	MYKEMYEV.1	2,192	"	30%
3	MYKEMYEV.2	1,988	"	27%
3	MYLU.0	4,194	13,080	32%
3	MYLU.1	3,399	"	26%
3	MYSE.0	1,464	2,139	68%
3	MYSO.0	627	1,577	40%
3	MYSO.1	502	"	32%
3	MYTH.0	832	848	98%
3	MYVO.0	308	1,258	24%
3	MYVO.1	357	"	28%
3	MYYU.0	19,947	23,010	87%
3	NYHU.0	1,196	2,251	53%
3	NYHU.1	615	"	27%
3	PESU.0	1,298	1,504	86%
		82,432	111,762	74%

A feature vector is generated for each call by combining the Fisher scores for all of the models and their prior probabilities.

For each pair of clusters across all species, a balanced Winnow classifier is trained to find an optimal hyper-plane separating each pair. Additional Winnow classifiers are trained between each cluster and all discarded clusters which are labeled as "no identification". On average, the Winnow classifiers successfully separated 99.95%, 99.06% and 97.17% of call pairs for the first, second and third dataset respectively.

For testing, a feature vector is extracted from each call of each test file as above. The feature vector is then classified by each of the binary Winnow classifiers and a hinge loss function is utilized to determine the overall winning class with a confidence level indicated by the combined classifier margins. If the "no identification" label wins, then the call is discarded. At the file level, the confidence of each winning call is combined to determine an overall winning species classification with an overall confidence score.

Receiver Operator Characteristic (ROC) curves are calculated and normalized to account for the different distributions of species in the datasets. As illustrated in the figures below, a ROC curve is a graph of the False Positive Rate (FPR, or Type I error rate) vs. the True Positive Rate (sensitivity, or the rate of actual negatives less Type II errors) at different confidence thresholds for a given classifier. Operating a classifier at a higher threshold means that classifications with lower confidence will be discarded which generally improves the accuracy of the classifications by reducing the False Positive Rate. However, this comes at the expense of increasing the False Negative Rate (Type II errors). The classifier thresholds can therefore be adjusted to meet the performance requirements of the application.

ROC curves are usually plotted in the unit square, but are shown here with the False Positive Rate axis only out to 0.10 to show more detail in the results. The perfect classifier runs along the vertical axis and perfect classification occurs in the upper left corner of the square. The "line of no discrimination" is the diagonal from the origin to the upper right corner of the unit square. Classifiers falling below this line are no better than a random coin toss.

The thresholds are optimized to strike a balance between system-wide misclassification rates and false negative rates by assigning a cost to each of these two types of errors and choosing thresholds to minimize this cost. For identifying bats from field recordings, it may be more important to optimize classifier performance for accuracy rather than sensitivity. Inaccurate classifications can have a high cost, for example incorrectly detecting the presence of endangered bats when no such bats are present. On the other hand, reduced sensitivity can be easily overcome by monitoring for bats over longer periods of time to increase the chances of detection and thus accurate classification.

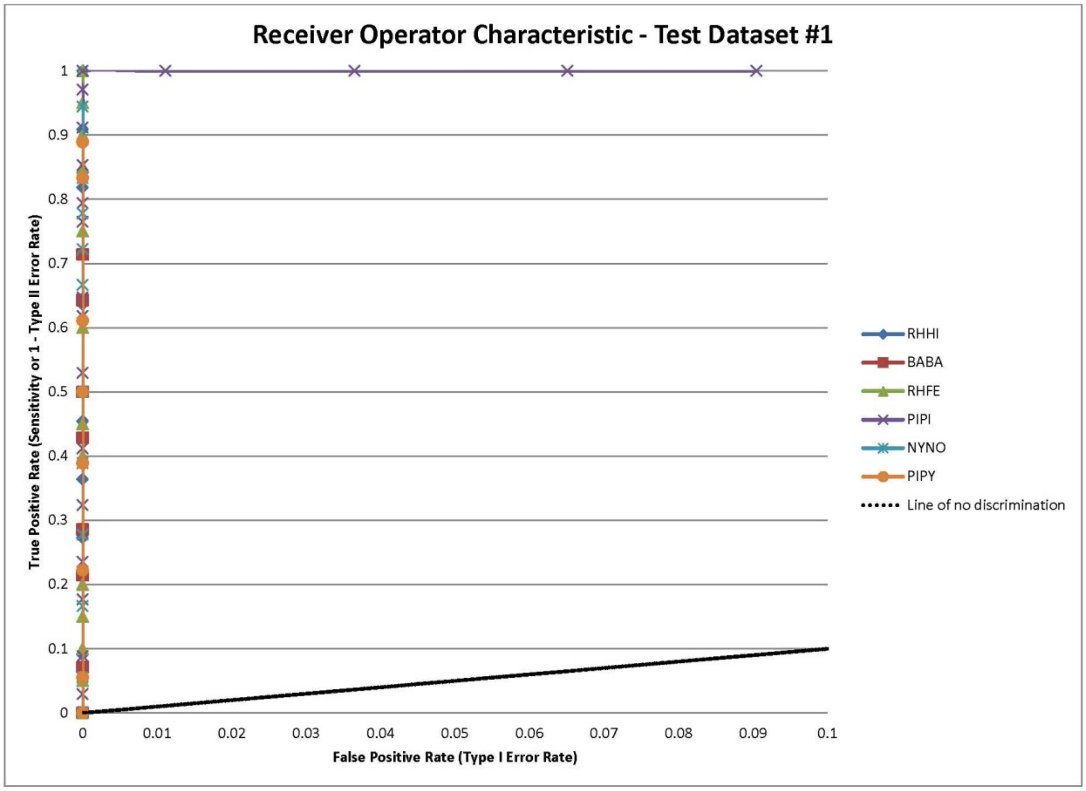

Figure 6: Receiver Operator Characteristic for Dataset #1

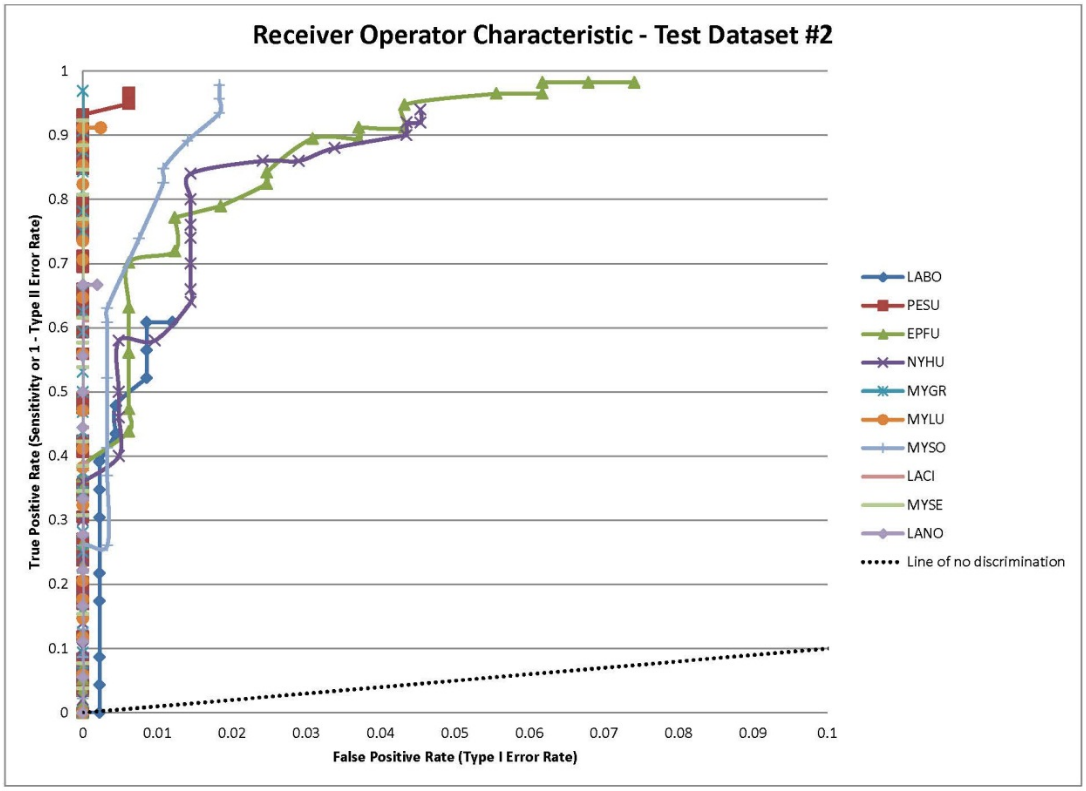

Figure 7: Receiver Operator Characteristic for Dataset #2

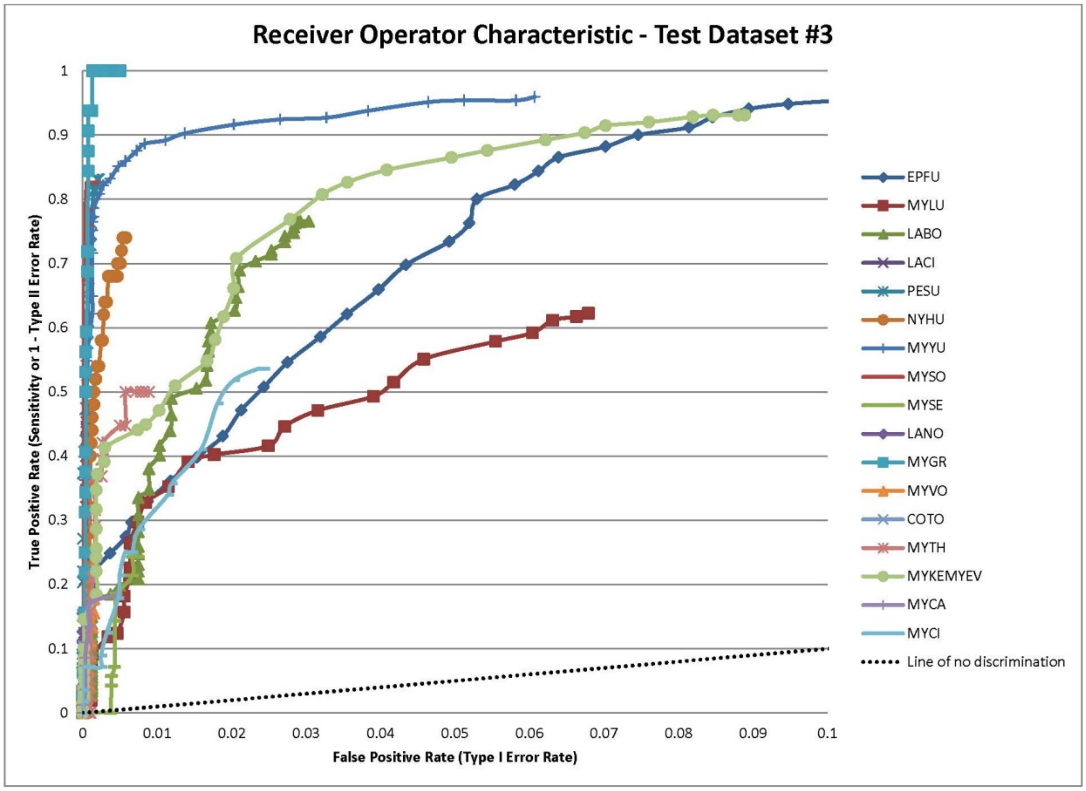

Figure 8: Receiver Operator Characteristic for Dataset #3

The thresholds are optimized to favor accuracy over sensitivity. More specifically, the cost of a misclassification was set to ten times the cost of a false negative. It is important to note that this cost preference can be customized to the application and that higher sensitivity levels can be achieved at the expense of lower accuracy.

The confusion matrixes are shown in the tables below for both training and testing data for all datasets using these optimized thresholds. Each row represents actual recordings with species labels and each column represents a classification result. The Correct Classification Rate (CCR) is the number of correct classifications divided by the total number of classifications for a given input. The Positive Predictive Value (PPV) is a measure of classifier performance and is the number of correct classifications divided by the total number of classifications for a given classifier. The True Positive Rate (TPR), or sensitivity, is the number of correct classifications divided by the total number of corresponding inputs for a given classifier and is inversely related to Type II errors. The False Positive Rate (FPR) is the number of incorrect classifications divided by the total number of classifications for a given classifier and is related to Type I errors. Note that PPV, TPR, and FPR are normalized to account for the differences in the distribution of species in the datasets.

Table 5: Confusion Matrix - Training Data (Dataset #1)

	RHHI	BABA	RHFE	PIPI	NYNO	PIPY	CCR	FILES	CALLS
RHHI	**11**						100.0%	11	77
BABA		**13**					100.0%	14	73
RHFE			**20**				100.0%	20	131
PIPI				**35**			100.0%	35	840
NYNO					**18**		100.0%	19	300
PIPY						**18**	100.0%	18	379
						Correct Classification Rate	100.0%	117	1,800

PPV	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	*Positive Predictive*
TPR	100.0%	92.9%	100.0%	100.0%	94.7%	100.0%	97.9%	*True Positive Rate*
FPR	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	*False Positive Rate*

Table 6: Confusion Matrix - Test Data (Dataset #1)

	RHHI	BABA	RHFE	PIPI	NYNO	PIPY	CCR	FILES	CALLS
RHHI	**11**						100.0%	11	68
BABA		**10**					100.0%	14	85
RHFE			**20**				100.0%	20	141
PIPI				**34**			100.0%	34	622
NYNO					**17**		100.0%	18	251
PIPY						**16**	100.0%	18	342
						Correct Classification Rate	100.0%	115	1,509

PPV	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	*Positive Predictive*
TPR	100.0%	71.4%	100.0%	100.0%	94.4%	88.9%	92.5%	*True Positive Rate*
FPR	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	*False Positive Rate*

The first dataset consists of a relatively small number of files and species. These species are generally considered easy to tell apart, and it is entirely possible that some of the test data came from the same individual bats as the training data given the small sample size. It is therefore not surprising that the classification performance on this dataset was near perfect. There were no misclassified files and 92.5% of all test files were identified to species.

Table 7: Confusion Matrix - Training Data (Dataset #2)

	LABO	PESU	EPFU	NYHU	MYGR	MYLU	MYSO	LACI	MYSE	LANO	CCR	FILES	CALLS
LABO	**23**										100.0%	24	923
PESU		**58**									100.0%	60	1,504
EPFU			**55**								100.0%	57	2,097
NYHU				**46**							100.0%	51	2,251
MYGR					**32**						100.0%	33	1,453
MYLU						**34**					100.0%	34	1,168
MYSO							**47**				100.0%	47	1,577
LACI								**17**			100.0%	18	422
MYSE									**27**		100.0%	27	1,195
LANO										**18**	100.0%	18	425
										Correct Classification Rate	100.0%	369	13,015

PPV	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	*Positive Predictive*
TPR	95.8%	96.7%	96.5%	90.2%	97.0%	100.0%	100.0%	94.4%	100.0%	100.0%	97.1%	*True Positive Rate*
FPR	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	*False Positive Rate*

Table 8: Confusion Matrix - Test Data (Dataset #2)

	LABO	PESU	EPFU	NYHU	MYGR	MYLU	MYSO	LACI	MYSE	LANO	CCR	FILES	CALLS
LABO	**14**			1							93.3%	23	815
PESU	1	**55**									98.2%	59	1,717
EPFU			**22**								100.0%	57	1,889
NYHU	3			**29**							90.6%	50	2,120
MYGR					**31**						100.0%	32	1,382
MYLU						**31**	1				96.9%	34	1,226
MYSO							**29**				100.0%	46	1,559
LACI								**11**			100.0%	18	380
MYSE									**24**		100.0%	26	707
LANO										**12**	100.0%	18	349
										Correct Classification Rate	97.9%	363	12,144

PPV	88.8%	100.0%	100.0%	93.0%	100.0%	100.0%	95.5%	100.0%	100.0%	100.0%	97.7%	*Positive Predictive*
TPR	60.9%	93.2%	38.6%	58.0%	96.9%	91.2%	63.0%	61.1%	92.3%	66.7%	72.2%	True Positive Rate
FPR	0.9%	0.0%	0.0%	0.5%	0.0%	0.0%	0.3%	0.0%	0.0%	0.0%	0.2%	False Positive Rate

The second dataset also performed extremely well with only 6 test files misclassified out of 363. This is a larger dataset with 10 species of bats, and some are considered easily confused including the MYLU/MYSO pairing discussed earlier. Without further analysis, it seems reasonable to assume that the high classification rate is due to the fact that the files in this dataset are all of very high quality selected for containing only search phase calls. There would not be a large variety of random noise, clutter calls, feeding buzzes, and other confusing signals that are more common in the field. The average correct classification rate was 97.9% and an average of 72.2% of all files were identified to species.

Table 9: Confusion Matrix - Training Data (Dataset #3)

	EPFU	MYLU	LABO	LACI	PESU	NYHU	MYYU	MYSO	MYSE	LANO	MYGR	MYVO	COTO	MYTH	MYKE MYEV	MYCA	MYCI	CCR	FILES	CALLS
EPFU	**537**			6					3	1								98.2%	1,672	38,036
MYLU	1	**206**	3		1				3		2	1			1			94.5%	366	13,080
LABO	3	10	**533**		12	16	1		4		5							91.3%	778	9,334
LACI	2			**152**										1				98.1%	265	3,940
PESU					**58**													100.0%	60	1,504
NYHU						**44**												100.0%	51	2,251
MYYU		1					**267**							1		1		98.9%	370	23,010
MYSO								**46**										100.0%	47	1,577
MYSE									**51**								1	98.1%	71	2,139
LANO										**41**								100.0%	72	1,521
MYGR											**31**							100.0%	33	1,453
MYVO												**39**						100.0%	45	1,258
COTO													**21**					100.0%	29	451
MYTH														**33**				100.0%	36	848
MYKEMYEV									1						**132**		1	98.5%	363	7,335
MYCA									4							**49**		92.5%	81	2,119
MYCI		1															**36**	97.3%	57	1,906
																	Correct Classification Rate	98.1%	4396	111,762
PPV	95.8%	94.4%	98.8%	99.4%	98.2%	97.7%	99.8%	100.0%	91.4%	99.9%	98.8%	99.7%	100.0%	99.3%	99.3%	99.6%	97.4%	98.2%	Positive Predictive	
TPR	32.1%	56.3%	68.5%	57.4%	96.7%	86.3%	72.2%	97.9%	71.8%	56.9%	93.9%	86.7%	72.4%	91.7%	36.4%	60.5%	63.2%	70.6%	True Positive Rate	
FPR	0.1%	0.2%	0.1%	0.0%	0.1%	0.1%	0.0%	0.0%	0.4%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	False Positive Rate	

Table 10: Confusion Matrix - Test Data (Dataset #3)

	EPFU	MYLU	LABO	LACI	PESU	NYHU	MYYU	MYSO	MYSE	LANO	MYGR	MYVO	COTO	MYTH	MYKE MYEV	MYCA	MYCI	CCR	FILES	CALLS
EPFU	**372**			15					5	1								94.7%	1,671	38,525
MYLU		**119**	5				2					4		1	1	3	2	86.9%	363	13,482
LABO	4	11	**380**		11	43			5		10	1						81.7%	776	9,058
LACI	2			**120**										1				97.6%	265	3,727
PESU			1		**43**													97.7%	59	1,717
NYHU			8			**34**												81.0%	50	2,120
MYYU							**203**									3		98.5%	370	22,542
MYSO								**31**										100.0%	46	1,559
MYSE		1							**13**					1				86.7%	70	1,856
LANO	1									**11**								91.7%	71	1,482
MYGR											**30**							100.0%	32	1,382
MYVO												**8**						100.0%	45	1,138
COTO		1											**3**					75.0%	29	515
MYTH									1					**14**	1			87.5%	38	907
MYKEMYEV							1		2				1	2	**126**			95.5%	363	7,725
MYCA		3							1			1				**14**	6	56.0%	82	1,915
MYCI		2							1								**12**	80.0%	56	1,711
																	Correct Classification Rate	88.8%	4386	111,361
PPV	89.3%	70.8%	72.0%	98.1%	98.1%	92.5%	98.5%	100.0%	72.3%	99.6%	98.6%	87.9%	97.4%	93.3%	92.3%	91.2%	73.1%	89.7%	*Positive Predictive*	
TPR	22.3%	32.8%	49.0%	45.3%	72.9%	68.0%	54.9%	67.4%	18.6%	15.5%	93.8%	17.8%	10.3%	36.8%	34.7%	17.1%	21.4%	39.9%	True Positive Rate	
FPR	0.2%	0.8%	1.2%	0.1%	0.1%	0.3%	0.1%	0.0%	0.4%	0.0%	0.1%	0.2%	0.0%	0.2%	0.2%	0.1%	0.5%	0.3%	False Positive Rate	

The third dataset is more of a real world test given 17 species of bats recorded in a large variety of field conditions. Average correct classification rate was 88.8% with an average of 39.9% of all files identified to species. Most classifiers did very well, but some had higher classification errors, most notably *Corynorhinus townsendii* (COTO). As mentioned earlier, COTO is the most underrepresented in the training data with only 29 training files. The performance of classifiers on the training data is significantly better compared to the test data. For training data, the average correct classification rate was 98.1%

with 70.6% of all files identified to species. This suggests that additional training data may improve classification performance.

The most important result is that two species of U.S. endangered bats including *Myotis sodalis* (MYSO) and *Myotis grisescens* (MYGR) are accurately detected and identified. The MYSO classifier has a Correct Classification Rate of 100%, a False Positive Rate of 0.0%, and a True Positive Rate of 67.4% meaning that approximately two in every three bat passes will be detected, and once detected, will be accurately identified. The MYGR classifier has a Correct Classification Rate of 100%, a False Positive Rate of 0.1%, and a True Positive Rate of 93.8% meaning nearly all bat passes will be detected and identified correctly. These results suggest that the methods described are well suited to the accurate detection of these protected species.

Future work

Up to this point, the author has avoided review of the underlying recordings in any detail. In fact, the strength of these methods is the ability to build classifiers from raw data without further human analysis. One obvious next step is to study in greater detail common misclassifications to better understand how the methods can be improved further. Additionally, analysis of clustering and HMM parameters may offer insights into the structure of the echolocation calls of bats. We are also interested in applying these classification methods to other animal vocalizations such as birds, frogs and cetaceans.

Acknowledgements

I would like to thank the following people and organizations for making bat call recordings available for our research:

Dr. Cori Lausen, Birchdale Ecological Ltd., B.C. Canada

Lynn Robbins, Professor, Dept. of Biology, Missouri State University

Paul Howden-Leach, Baker Consultants Ltd., Derbyshire, United Kingdom

U.S. National Park Service

Bibliography

Agranat I. Automatically Identifying Animal Species from their Vocalizations [Report]. - Concord : Wildlife Acoustics, Inc., 2009.

Allen, Burt, and Miller Environmental Effects on the Echolocation Call Structure of Bats [Report]. - [s.l.] : Truman State University NSF-STEP Program, 2007.

Allen, C.R., S.E. Romeling, and L.W. Robbins Acoustic Monitoring and Sampling Technology [Conference] // Proceedings of Protected threatened bats at coal mines: a technical interactive forum. - [s.l.] : U.S. Dept. of Interior, 2011. - Vols. 173-188.

Aran and Akarun A Multi-class Classification Strategy for Fisher Scores: Application to Signer Independent Sign Language Recognition [Report]. - Martigny : Idiap Research Institute, 2009.

Barclay R. M. Bats are not birds - a cautionary note on using echolocation calls to identify bats: a comment [Journal] // Journal of Mammalogy. - 1999. - 1 : Vol. 80. - pp. 290-296.

Britzke, E.R., J. Duchamp, R.S. Swhiart, K.M. Murray, and L.W. Robbins Acoustic identification of bats in the eastern United States: A comparison of parametric and nonparametric methods [Journal]. - [s.l.] : Journal of Wildlife Mangement, 2011. - Vols. 75:660-667.

Butler M. Hidden Markov Model Clustering of Acoustic Data [Report] / School of Informatics ; University of Edinburgh. - [s.l.] : University of Edinburgh, 2003.

Chen, Man, and Nefian Face recognition based on multi-class mapping of Fisher Scores [Journal]. - [s.l.] : The Journal of the Pattern Recognition Society, 2005. - Vol. 38.

Corcoran A. Automated Acoustic Identification of Nine Bat Species of the Eastern United States [Report]. - [s.l.] : Humboldt State University, 2007.

Fukui, Agetsuma, and Hill Acoustic Identification of Eight Species of Bat (Mammalia Chiroptera) Inhabiting Forests of Southern Hokkaido, Japan: Potential for Conservation Monitoring [Journal]. - [s.l.] : Zoological Society of Japan, 2004. - Vol. 21. - pp. 947-955.

Kivinen and Warmuth The Perceptron algorithm vs. Winnow: linear vs. logratithmic mistake bounds when few input variables are relevant [Report] / Baskin Center for Computer Engineering & Information Sciences. - Santa Cruz : University of California, 1995.

Parsons and Jones Acoustic Identification of Twelve Species of Echolocating Bat by Discriminant Function Analysis and Artificial Neural Networks [Report]. - Bristol : School of Biological Sciences, University of Bristol, 2000.

Redgwell, Szewczak, Jones, and Parsons Classification of Echolocation Calls from 14 Species of Bat by Support Vector Machines and Ensembles of Neural Networks [Journal] // Algorithms. - 2009. - Vol. 2. - pp. 907-924.

Skowronski and Harris Bat detection/classification using machine learning [Journal] // J. Acoust. Soc. Am.. - [s.l.] : IEEE, 2006. - 3 : Vol. 119. - pp. 1817-1833.

Szewczak J. Acoustic ambiguity of Myotis lucifugus and M. sodalis [Conference] // North Eastern Bat Working Group 2012 Annual Meeting. - Carlisle, PA : [s.n.], 2011.