AUTOMATICALLY IDENTIFYING ANIMAL SPECIES FROM THEIR VOCALIZATIONS

Ian Agranat Wildlife Acoustics, Inc., Concord, Massachusetts, USA

ABSTRACT

Commercially available autonomous recorders for monitoring vocal wildlife populations such as birds and frogs now make it possible to collect thousands of hours of audio data in a field season. Given limited resources, it is not practical to manually review this volume of data “by ear”. The automatic processing of sound recordings to detect and identify specific species from their vocalizations, even if not perfectly accurate, makes efficient use of researchers who review only those samples most likely to contain vocalizations of interest. This results in significant gains of sample coverage, operating efficiency, and cost savings.

Developing generalized computer algorithms capable of accurate species identification in real-world field conditions is full of difficult challenges. First, recordings made by autonomous recorders typically receive sounds from all directions, scattered and reflected by trees, obscured by an unpredictable constellation of random noise, wind, rustling leaves, airplanes, road traffic, and other species of birds, frogs, insects and mammals. Second, the vocalizations of many species are highly varied from one individual to the next. Any algorithm must be prepared to accept vocalizations that are similar, but not identical, to known references in order to successfully detect the previously unobserved individual. However, in so doing, the algorithm is then susceptible to misclassifying a vocalization from a different species with similar components. This is especially true for species with narrowband vocalizations lacking distinctive spectral properties and in species with short duration vocalizations lacking distinctive temporal properties.

The bulk of prior research has generally differentiated among only a handful of simple mono-syllabic vocalizations at a time. While the results have been promising, we found that many approaches degrade significantly as the number of species increases, especially when more complex multi-syllabic and highly variable vocalizations are also included.

In this paper, we discuss an algorithm based on Hidden Markov Models automatically constructed so as to consider not just the spectral and temporal features of individual syllables, but also how syllables are organized into more complex songs. Additionally, several techniques are employed to reduce the effects of noise present in recordings made by autonomous recorders.

.

1 INTRODUCTION

Wildlife Acoustics was founded in 2003 to develop hardware and software capable of real-time field identification of birds from their songs. Initial research and experimentation used commercially available bird song recordings (e.g. from audio CDs available in bookstores). However, it was quickly apparent that the significant variation among individuals typical in many species would require substantially greater amounts of data to train and test different algorithms.

We acquired over 1,800 individual recordings spanning over 73 hours of 168 species common to North America from the Macaulay Library at the Cornell Lab of Ornithology.

For results published later in this paper, we are focused on a subset consisting of 52 species common to New England (North Eastern United States) including 2 distinctly different vocalizations of the Black-capped Chickadee and the Northern Flicker for a total of 54 vocalization classes. This subset is represented by 550 individual recordings manually segmented into 12,653 vocalization instances.

The algorithms described in this paper are embodied by our Song Scope application software (version 2.3). These algorithms have also been implemented on a proprietary DSP platform capable of real-time classification.

The scope of this paper is limited to describing our algorithms generally and demonstrating their classification performance across a wide range of vocalizations including several species with large repertoires and significant individual variation (e.g. Song Sparrow, Carolina Wren, etc.), several with narrowband whistled vocalizations (e.g. Northern Cardinal, Tufted Titmouse, etc.), several with broadband vocalizations (e.g. American Crow, Canada Goose, etc.), and several pairings that are commonly confused by human listeners (e.g. Purple Finch vs. House Finch, Yellow Warbler vs. Chestnut-Sided Warbler, etc.) with limited training data. Comparisons to other classification techniques are beyond the scope of this paper. While the testing methodology discussed does not address signal detection or performance in analyzing recordings made by autonomous recorders, our algorithms have been proven effective. Our previous publication [7] describes use of these algorithms to detect Cerulean Warblers from autonomous field recordings collected in several sites of the Allegheny National Forest. In addition, many of our customers have been using these algorithms successfully to detect a number of bird and frog species, and expect several to publish their results in the near future.

2 ALGORITHM

2.1 BACKGROUND

The Song Scope classification algorithms are based on Hidden Markov Models (HMMs) using spectral feature vectors similar to Mel Frequency Cepstral Coefficients (MFCCs) as these methods have proven effective in robust speech recognition applications. [1]

However, classical speech recognition algorithms are not well suited to the task of animal species classification in real-world field conditions for a number of reasons:

First, human speech recognition algorithms are advantaged by the speaker talking directly into a microphone such as a telephone receiver. In real-world field conditions, the animal vocalization of interest may be coming from 10-100 meters away, scattered and reflected by trees, obfuscated by the constellation of other sounds picked up by an omni-directional field microphone such as those from other species of birds, frogs, insects, and mammals; rustling leaves from wind; dripping water; and man-made sounds from automobiles, airplanes, and other sources. As a consequence, the signal-to-noise ratio (SNR) of vocalizations on field recordings is relatively weak. Even if distinctive spectral features are present in the vocalization, they may not be detectable above the noise floor.

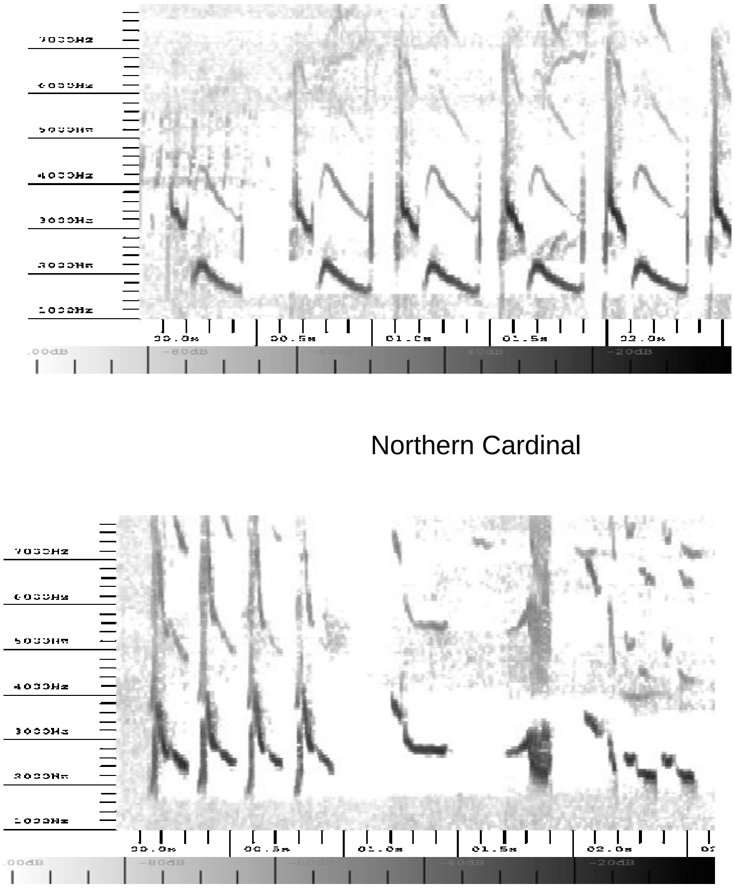
Second, human speech contains unique broadband spectral properties concentrated in a range of frequencies below approximately 3kHz. But animal vocalizations come in many different varieties. Many birds and some frogs have vocalizations consisting of narrowband whistled vocalizations lacking any distinctive spectral features. Others have broadband vocalizations with complex spectral properties. The vocalizations may occur in a wide range of frequencies from 100Hz to 10,000Hz. Some vocalizations are long, lasting several seconds, while others last only a fraction of a second. It is challenging to develop a single algorithm capable of accurate classification across such a broad range of vocalization types.

Third, human speech research has the benefit of large libraries of training data available to cover the range of individual variation across thousands of individual speakers. However, our resources are limited to an average of ten individuals of each species making it difficult to model the wide range of variation typical of many species. There are also significant acoustic differences between training data available, typically from recordings made with high quality parabolic microphones, and field data from omni-directional microphones captured by autonomous recorders.

Much of the prior research, particularly with respect to birdsong classification, has demonstrated various techniques such as Dynamic Time Warping, Spectrographic Cross-Correlation, traditional HMMs and Artificial Neural Networks to match the patterns of specific syllables of vocalizations across a narrow range of species. [3, 4, 5, 6] However, we found that syllable matching does not scale to a large number of species, especially when several highly variable narrowband vocalizations are included in the mix. Given individual variation, any algorithm must be generous in what it accepts as a match in order to detect the previously unobserved vocalization. However, in so doing, the algorithm becomes susceptible to false positives from other species with vocalizations containing similar syllables. This is especially true of narrowband whistled vocalizations lacking distinctive spectrographic features.

By way of example, consider the Northern Cardinal and the Baltimore Oriole (Figure 1). Both of these species have narrowband whistled vocalizations in the same general range of frequencies, (fundamentals 1 to 4 kHz) and both species are subject to significant individual variation. A whistled syllable of a Northern Cardinal may easily be confused with a whistled syllable from a Baltimore Oriole. However, a trained human listener can tell these two species apart by considering the entire vocalization, not by listening to just one syllable. Northern Cardinals tend to combine several similar up-slurred or down-slurred whistles compared to the Baltimore Oriole’s more varied and complex songs.

Song Scope algorithms combine several noise reduction techniques to enhance weak signals typical in field recordings and normalize them to match higher quality recordings available for training data. Next, signal detection techniques are employed to isolate individual vocalizations. Vocalizations are then transformed into a time-series of spectral feature vectors for analysis. Finally, Hidden Markov Models are built to model both the spectral and temporal features of individual syllables as well as the syntax of how syllables are combined to form more complex song. The Viterbi algorithm [1] is used to determine the statistical fit of candidate detections with the model while additional statistical filters are applied to further reduce false positives.



Northern Cardinal

Baltimore Oriole

Figure 1: Spectrograms of Vocalizations with similar narrowband syllables. The spectrograms span approximately 2.5 s and the frequency range 1.0-8.0 kHz.

2.2 NOISE REDUCTION

The Song Scope algorithms first pre-process the audio signal using a number of techniques to reduce the effects of noise typical in recordings created by automatic remote monitoring systems.

Figure 2 shows a spectrogram of a Northern Cardinal recording captured by a Song Meter, our autonomous audio recorder. Notice that this spectrogram is “blurry”, a consequence of omni-directional microphones picking up the signal as it is scattered and reflected by the trees. Also notice the low frequency random noise produced by nearby automobile traffic. Finally, note the short vocalization from a different species visible in the upper left corner of the spectrogram.

As a side note, observe how different this Northern Cardinal vocalization is from that shown in Figure 1, further illustrating the variability among individuals typical in many species.

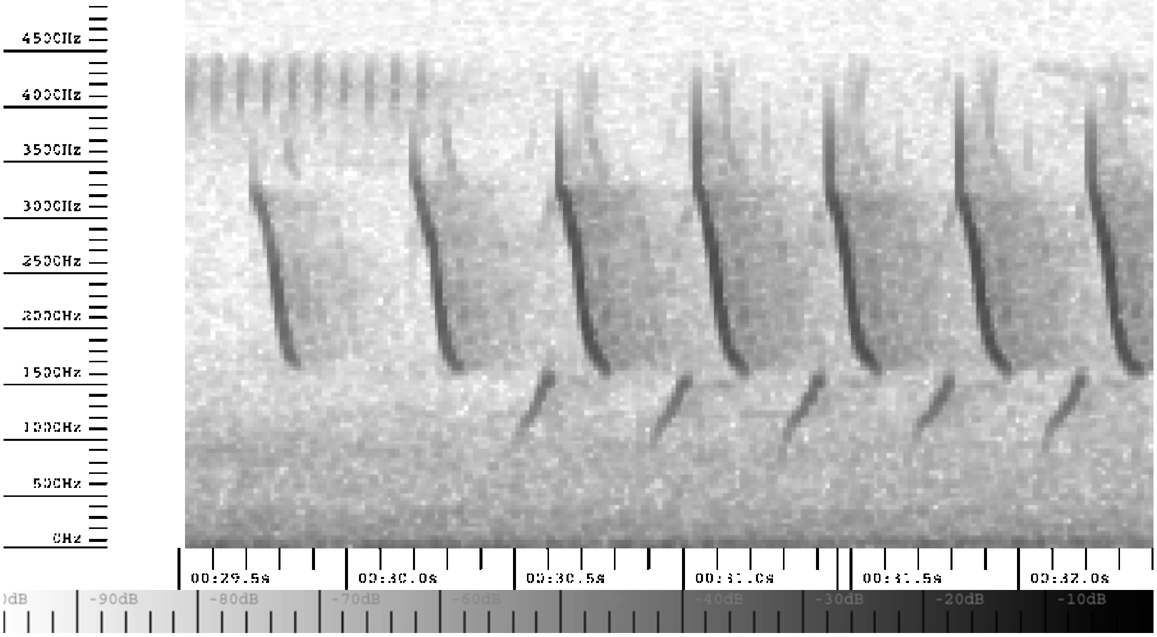


Figure 2: Spectrogram of Original Field Recording. The spectrogram spans approximately 3.0 s and the frequency range 0.0 – 5.0 kHz.

The first step in noise filtering is to apply a Wiener filter [2] to the signal to reduce stationary background noise. The Wiener filter requires an estimate of the noise spectrum, and Song Scope calculates this by using a simple one-second rolling average of the spectrum immediately preceding each FFT window. Figure 3 shows the filtered spectrogram. Notice how the blurring is almost completely eliminated and the low frequency background noise is reduced.

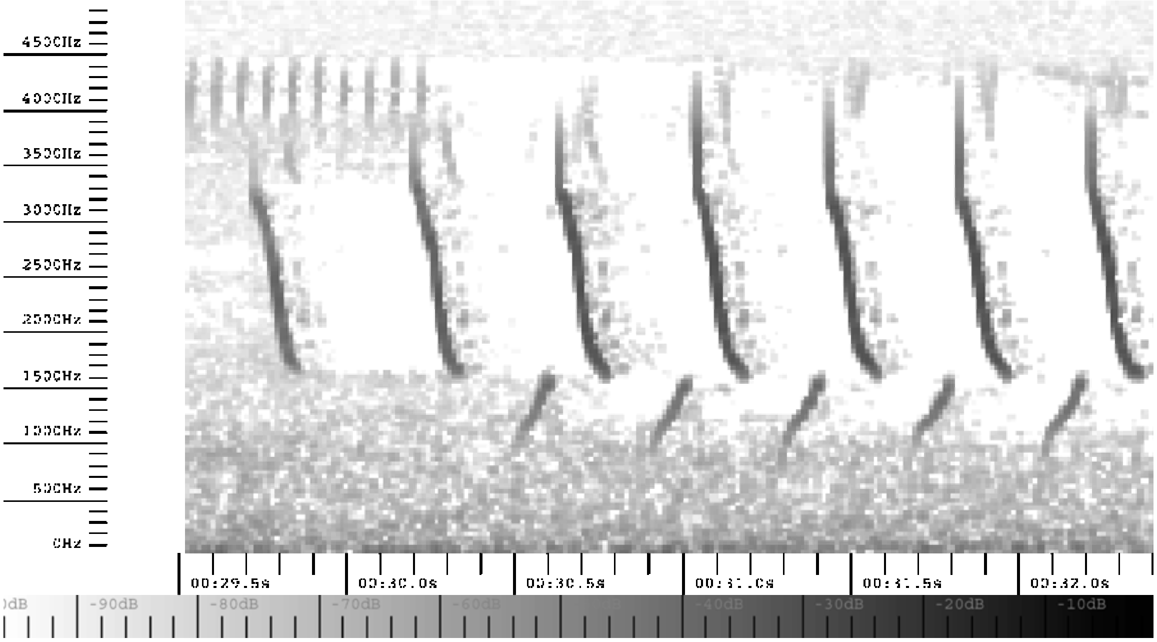


Figure 3: Spectrogram with Wiener Filter Applied

The second step in noise filtering is to apply a band-pass filter to consider only the range of frequencies that a target species is likely to produce. This helps reduce interference from other sounds that may occur at higher or lower frequencies, especially wind and traffic noise, which tend to be more intense at lower frequencies. (Some of our studies found that “pink noise” was a good estimate of background noise, where power is inversely proportional to frequency).

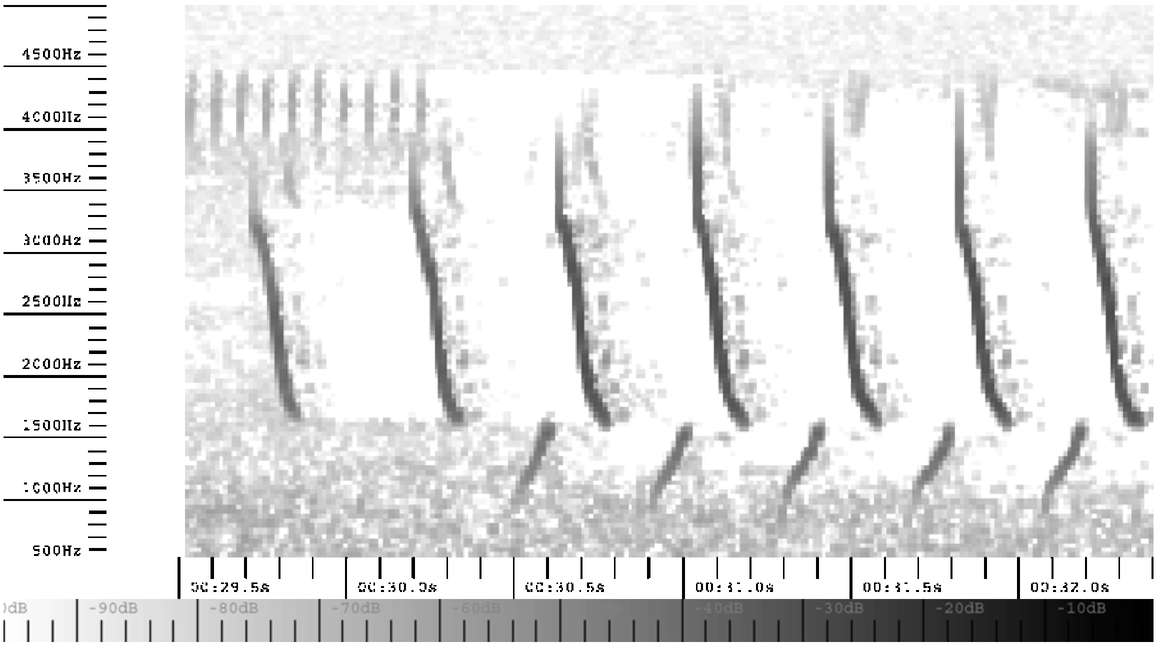


Figure 4: Band pass filter 500-5000Hz

Figure 4 illustrates a simple example, actually a high-pass filter at approximately 500Hz. The filter is implemented in the frequency domain by simply ignoring frequencies outside the selected band.

The next step is not directly related to noise reduction, but is applied to the signal at this point. Song Scope redistributes the spectrum from a linear frequency scale to a log frequency scale. This step is similar to Mel frequency transforms common in speech recognition. In speech, the Mel scale is a logarithmic frequency scale designed to represent the sensitivity of the human ear assuming important spectral features in speech are likely to correspond to the human ear's sensitivity to different frequencies.

We take a slightly different view. Rather than attempting to model the hearing sensitivity of various animal species, we use the log frequency scale transformation as a means of spectral feature reduction, under the assumption that higher frequency components are redundant harmonics of the fundamental vocalization frequencies. The log scale emphasizes the lower fundamental frequencies while deemphasizing higher frequency harmonics. The actual log scale is adapted to the specific band-pass filter according to Formula 1 where the *nth* log-spaced bin F_n is derived from the band-pass filter's minimum frequency f_a, and constant k. The value of k is chosen such that the number of log frequency bins is equal to the original number of linear frequency bins included in the band-pass filter. Note that the high-pass cut-off frequency represented by f_a should be as high as possible for a given vocalization to maximize the emphasis of fundamental frequency components.

$$F_n = f_a 2^{kn} \qquad (1)$$

We achieved higher classification performance with non-overlapping rectangular filter banks rather than the overlapping triangular filter banks used in MFCCs.

Figure 5 illustrates the log-warped frequency transformation.

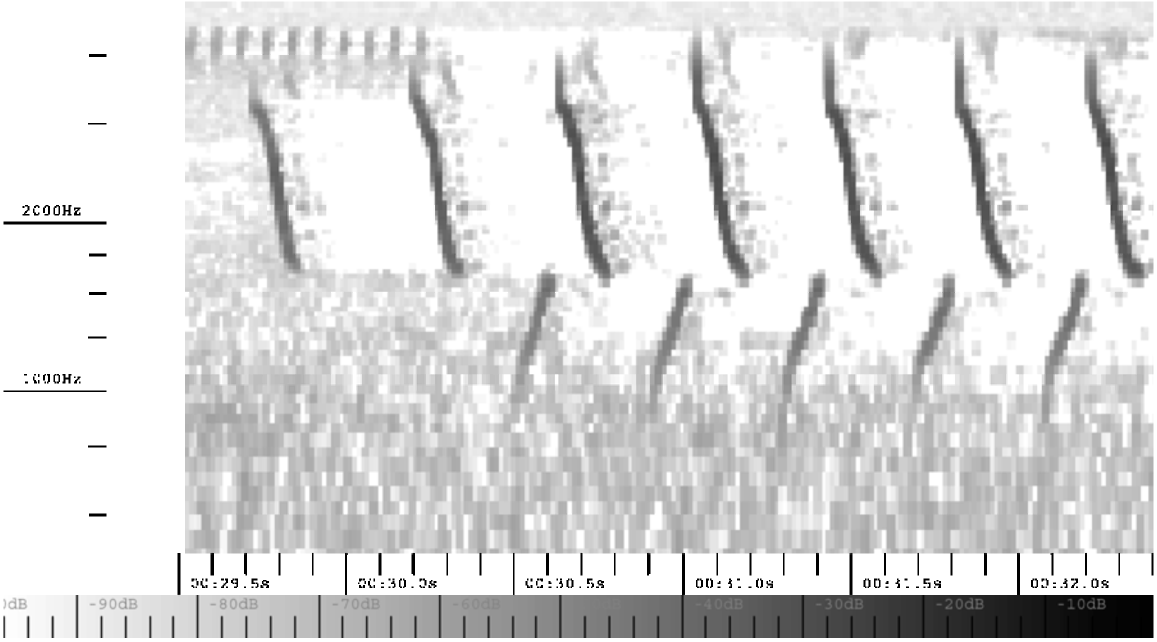


Figure 5: Log Frequency Transformation

Finally, log power levels are normalized according to a fixed dynamic range. The power level of each frequency bin for a given FFT window is shifted such that the frequency bin with the highest energy level is set to equal the dynamic range. Any bin whose normalized power falls below zero or below the estimated background noise level for that bin is set to zero. This results in a dramatic reduction of noise as illustrated in Figure 6.

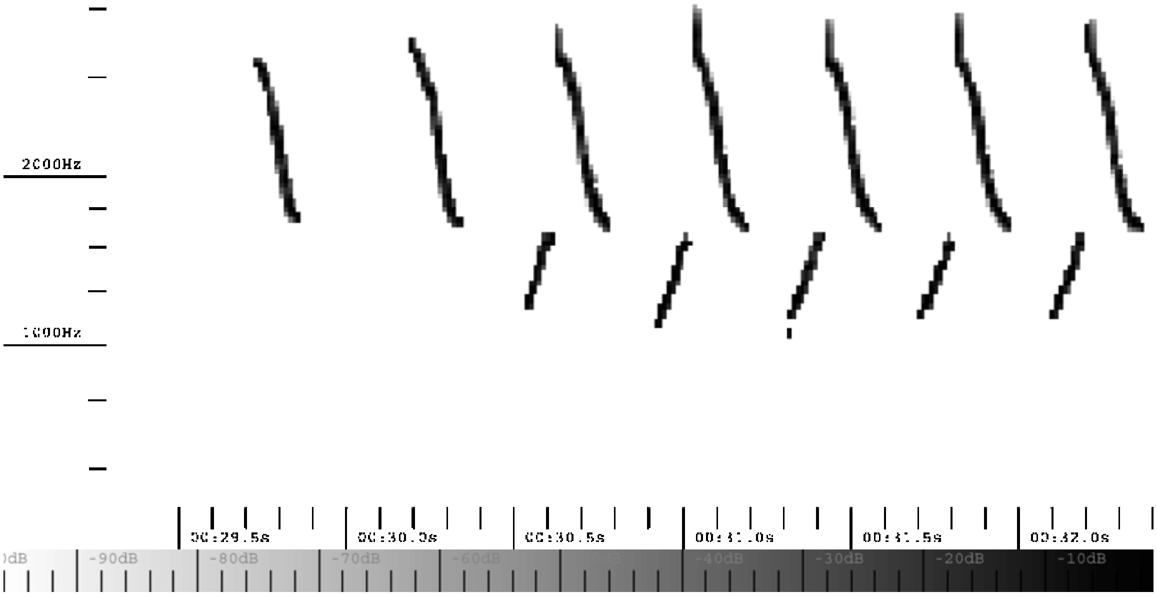


Figure 6: Power Normalization to a Fixed Dynamic Range

There is a trade-off here in that higher values of dynamic range will provide more spectral detail for classification but will also be more susceptible to background noise. We have found that selection of the dynamic range to reflect the expected signal-to-noise ratio of the vocalization in the field provides a realistic balance of spectral detail, typically 15-20dB. This also helps reduce detailed spectral information found in high quality training recordings made by parabolic microphones, to match the actual spectral detail present in noisier field recordings.

2.3 SIGNAL DETECTION

Song Scope uses a simple signal detection algorithm to locate the beginning and end of candidate vocalizations by monitoring the total energy passing through the band-pass filter. The following parameters are used to tune the signal detection algorithm for a particular vocalization:

Max Syllable Duration	- longest expected duration of any one syllable
Max Syllable Gap Duration	- longest expected gap between any two syllables
Max Song Duration	- longest expected vocalization
Dynamic Range	- normalized dynamic range (as discussed above)

A rolling window of length 2 x (*Max Syllable Duration* + *Max Syllable Gap Duration)* is monitored to track local minimum and maximum energy values. Except as noted below, a low-water mark is established at +12dB above the local minimum, and a high-water mark is established at +6dB above the low-water mark. The beginning of a syllable (onset) is defined when the energy level rises above the high-water mark and ends when the energy level falls below the low-water mark. If an inter-syllable gap is detected exceeding the *Max Syllable Gap Duration*, the vocalization is considered complete. The vocalization may also be considered complete during an inter-syllable gap when the *Max Song Duration* has been reached.

When an unusually strong vocalization is detected, the low-water and high-water marks are automatically adjusted upwards according to the *Dynamic Range* parameter. Specifically, the low-water mark is set to the local maximum energy level, less the *Dynamic Range*, less 3dB. The high-water mark is set to +6dB above the low-water mark as before. This helps normalize the signal detection to compensate for differences between high quality training recordings and low quality field recordings.

In addition, the strong onset of a vocalization may force a weaker background vocalization to terminate early, giving precedence to the stronger vocalization and an opportunity to analyze both signals independently.

Setting the *Max Syllable Gap Duration* parameter as small as possible helps the detector end one vocalization before the onset of the next. Otherwise, two different back-to-back vocalizations may be perceived as one.

2.4 FEATURE EXTRACTION

A series of Discrete Cosine Transform (DCT-II) coefficients are derived from the power-normalized and log-warped frequency bins. These are essentially MFCCs except that the frequency transformation is not specifically Mel scale as described earlier. A relative power level feature is also added. Finally, sequential features with essentially the same information may be coalesced into a single feature vector, with one additional component added to represent the feature duration.

2.5 HIDDEN MARKOV MODEL ESTIMATION

A Hidden Markov Model is typically a collection of states. Each state represents spectral properties in the form of Gaussian mixtures of spectral feature vectors, while temporal properties are represented by state transition probabilities.

Our objective is to model not just individual syllables within a vocalization, but the syntax of how many syllables might combine to form more complex vocalizations.

Song Scope automatically segments training data into individual syllables using cues from the signal detection algorithm and local signal power minima to find syllable boundaries. Syllables are then automatically clustered into classes of similar syllables. States are then assigned sequentially to each syllable class in proportion to the mean duration of syllables in the class. Some states represent the inter-syllable gaps. An initial Hidden Markov Model topology, Gaussian mixtures and transition probabilities can then be estimated. The model is then iteratively refined by finding the optimum state sequence using the well-known Viterbi algorithm, backtracking and recalculating model parameters until there is no further improvement. Figure 7 illustrates this process.

Note that Viterbi training is not the same as Baum-Welch training [1]. Viterbi training was chosen because it is computationally efficient and easily implemented, and is known to facilitate segmentation of syllable features across HMM states. However, Baum-Welch has generally outperformed Viterbi training in speech recognition applications. We hope to experiment with Baum-Welch training in the future to see if further improvement may be possible.

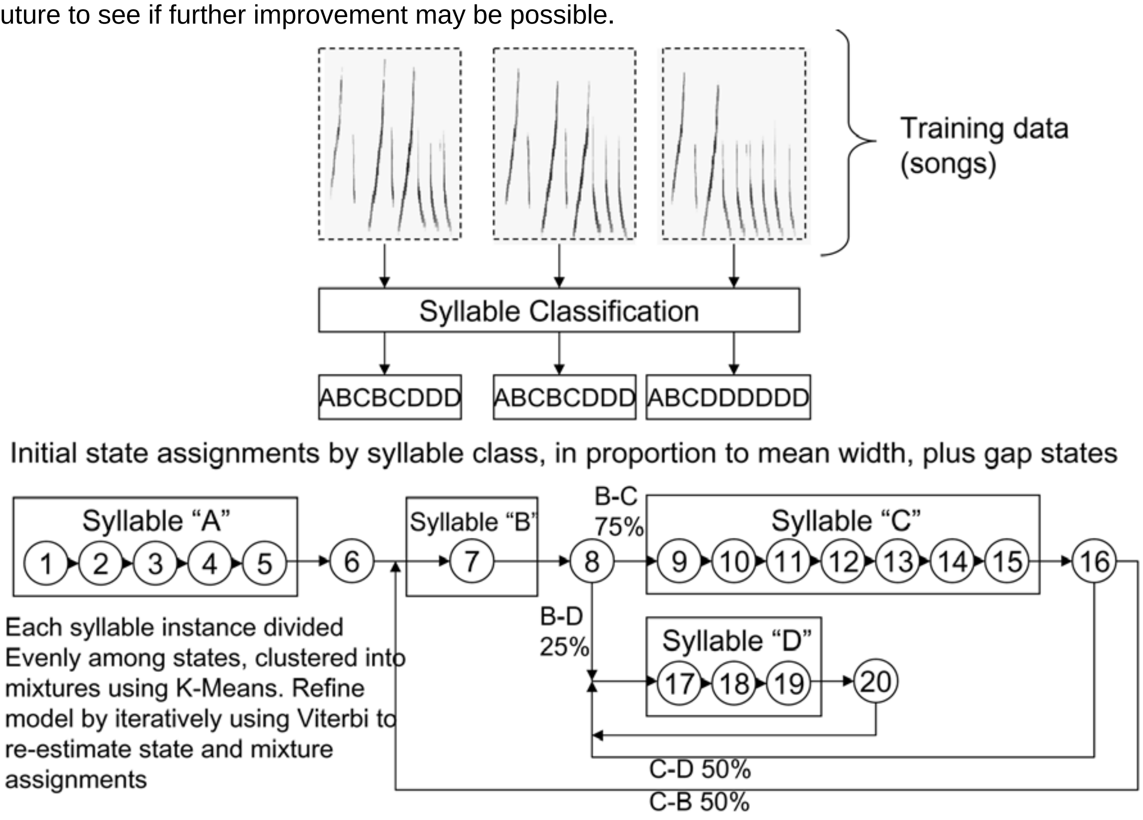

Training data
(songs)
Syllable Classification
ABCBCDDD
ABCBCDDD
ABCDDDDDD
Initial state assignments by syllable class, in proportion to mean width, plus gap states
Syllable "A"
1
2
3
4
5
6
Syllable "B"
7
B-C
75%
8
Syllable "C"
9
10
11
12
13
14
15
16
B-D
25%
Syllable "D"
17
18
19
20
Each syllable instance divided
Evenly among states, clustered into
mixtures using K-Means. Refine
model by iteratively using Viterbi to
re-estimate state and mixture
assignments
C-D 50%
C-B 50%

Figure 7: Hidden Markov Model Parameter Estimation

The automatic classification of syllables involves a two-dimensional DCT commonly used in image compression to reduce the temporal and spectral features of individual syllables into a feature vector. These features are then clustered using K-Means into individual syllable classes. Figure 8 illustrates this process.

Song Scope will build several different models by clustering syllables into a different number of classes. With too many classes, there is an increased risk of overtraining in that each training set may fall into its own class with a near perfect fit. To avoid over-training, each model is evaluated by scoring it against previously unobserved vocalizations in order to optimize the model to recognize previously unobserved data rather than maximizing the fit to training data. We accomplish this by excluding one of the training data recording sources from the model, and then testing the model against the excluded source. This process is repeated for each of the training sources for each candidate model. The model that scores highest against excluded data is then chosen and rebuilt with all training data included.

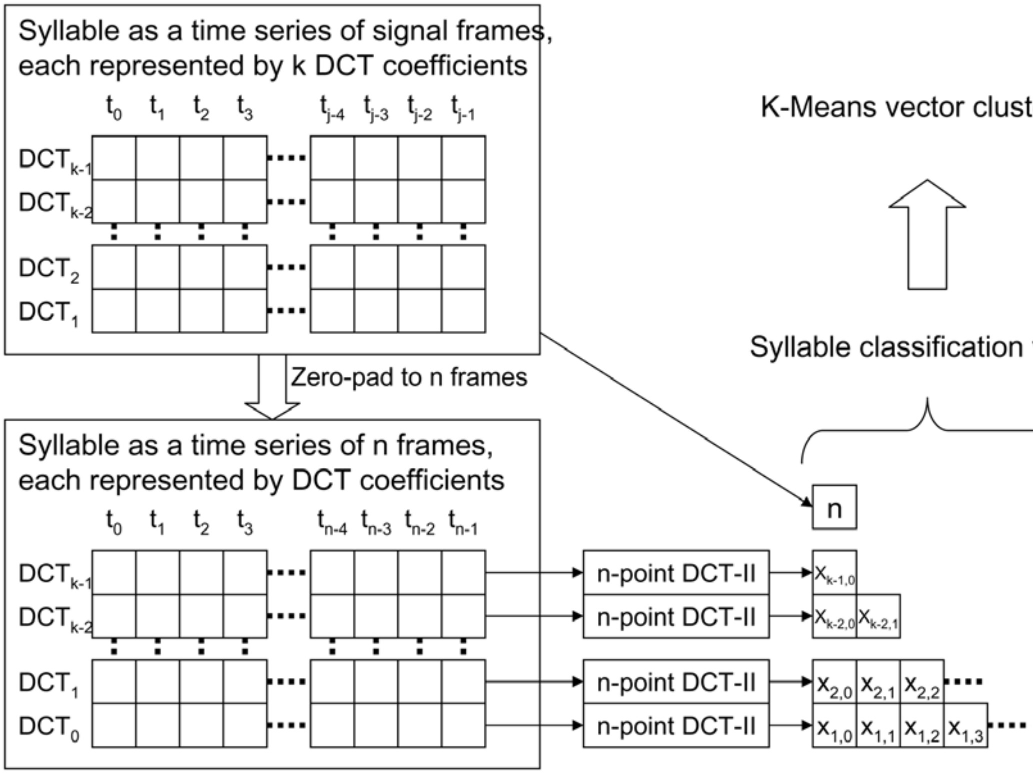

Syllable as a time series of signal frames, each represented by k DCT coefficients
t_0 t_1 t_2 t_3 t_{j-4} t_{j-3} t_{j-2} t_{j-1}
DCT_{k-1}
DCT_{k-2}
DCT_2
DCT_1
K-Means vector clustering
Syllable classification vector
Zero-pad to n frames
Syllable as a time series of n frames, each represented by DCT coefficients
t_0 t_1 t_2 t_3 t_{n-4} t_{n-3} t_{n-2} t_{n-1}
DCT_{k-1}
DCT_{k-2}
DCT_1
DCT_0
n-point DCT-II
n-point DCT-II
n-point DCT-II
n-point DCT-II
n
$x_{k-1,0}$
$x_{k-2,0}$ $x_{k-2,1}$
$x_{2,0}$ $x_{2,1}$ $x_{2,2}$
$x_{1,0}$ $x_{1,1}$ $x_{1,2}$ $x_{1,3}$

Figure 8: Clustering similar syllables into classes

2.6 ADDITIONAL STATISTICAL FILTERS

A goal of the Song Scope algorithm is to classify a wide range of vocalizations including narrowband and broadband with both short and long durations. One potential weakness of HMMs is that a short narrowband vocalization may be a close match to a Gaussian mixture within a larger model, resulting in a false positive. To compensate, a second layer of statistical filters is applied to compare candidate vocalizations to the training data. The *quality* value produced by Song Scope is a composite that includes the duration of the vocalization and the number of different HMM states traversed compared to the distribution observed in the training data. A combination of high *quality* score and HMM probability suggests a strong match.

By adjusting the minimum acceptable values for both the quality value and the HMM probability, the classifiers can be tuned to trade off between sensitivity and accuracy.

3 PERFORMANCE

3.1 TESTING METHODOLOGY

To measure the classification performance of Song Scope, we started by choosing 52 species of birds most common to New England. While many of these species have a variety of songs and calls among their vocalizations, we generally limited our testing to the more common song or call as we often only had one vocalization class represented in our collection of recordings. In two cases, the Black-capped Chickadee and the Northern Flicker, we also had a number of recordings of a distinct secondary vocalization, so these were included in the testing for a total of 54 vocalization classes.

We used 550 individual recordings acquired from the Macaulay Library at the Cornell Lab of Ornithology. On average, each recording lasts about 2.5 minutes and typically contains vocalizations of only one individual, although some recordings include many individuals.

The recordings were manually segmented into 12,653 vocalization instances and labeled as representing one of the 54 different vocalizations under study.

The 550 recordings were then split arbitrarily (by even/odd tracks on CDs) into 266 training recordings and 284 testing recordings. On average, each vocalization is represented by 5 training recordings and 5 testing recordings with 23 vocalization instances in each recording. This implies that most classifiers are built with vocalizations from only 5 individuals of each species.

The same set of Song Scope parameters was chosen for all 54 classifiers as follows:

Sample Rate	16,000 samples per second
FFT Window Size	256 samples (16ms)
FFT Window Overlap	50%
Band-pass filter	1kHz – 8kHz
Max Syllable Duration	1.5 seconds
Max Syllable Gap Duration	0.5 seconds
Max Song Duration	3.0 seconds
Dynamic Range	15dB
Max HMM Model States	48
HMM feature vector size	12
Mixtures per state	1

Note that better classification results may be possible if individual classifiers are tuned to maximize performance. However, using the same parameters allows us to run all classifiers in parallel across the testing data set with the best scoring model automatically tallied.

For each of the 6,384 test vocalization instances, Song Scope produces an HMM probability score and a *quality* value as described above. A minimum threshold for each of these two values was set at two standard deviations below the mean values observed on correct classifications. Once minimum thresholds were established, all results (approximately 7%) below these thresholds were discarded.

Note that the minimum thresholds allow for the balance between sensitivity and accuracy of detections. Low thresholds allow more detections thereby increasing sensitivity, but they may also allow more false positives resulting in reduced accuracy. High thresholds are more selective and reduce sensitivity, but may improve accuracy by considering only the strongest matches.

3.2 RESULTS

First we look at discrimination performance across the matrix of all 54 classifiers by considering only the highest scoring model that also achieves its minimum quality value and HMM probability thresholds.

For the training data, each classifier was able to detect an average of 63% of the target vocalizations with at least one vocalization detected on 95% of all target recordings. The false positive rate for each classifier on the training data was only 0.3%.

For the testing data, each classifier was able to detect on average 37% of the target vocalizations with at least one vocalization detected on 74% of all target recordings. The false positive rate for each classifier on the test data was only 0.4%.

The difference between testing and training data is to be expected as the HMMs and secondary statistical filters will fit the training data closely compared to the unobserved testing data. We would expect the testing data classification performance to improve with more training data as discussed below.

Figure 9 illustrates the classifier performance against unobserved test data by visualizing the matrix of all classifiers applied to all recordings on a three dimensional surface chart. Accurate detections are visible in the diagonal ridge where classifiers correspond to recordings. Gaps in the diagonal ridge indicate classifiers with low detection rates, and bumps in the surface outside the diagonal represent false positives. The most prominent false positive is the classifier for the Rose-breasted Grosbeak falsely detecting recordings of the American Robin. Those familiar with the vocalizations of these two species know that they are very similar and easily confused by untrained human observers.

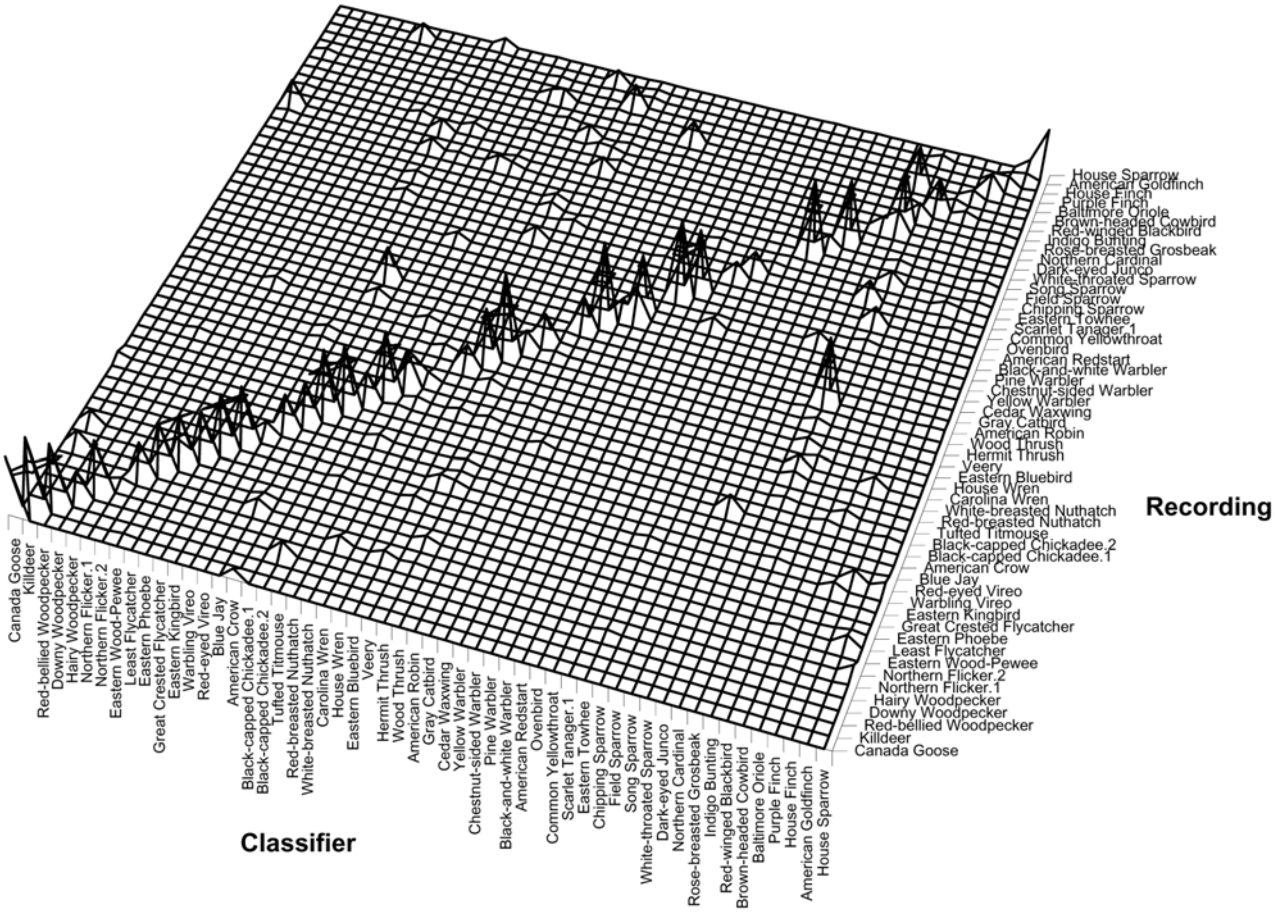


Figure 9: Classifier performance against unobserved test data

Next, we look at the performance of individual classifiers across the entire test data set. All matches that meet the minimum quality value and HMM probability are tallied without considering results from other classifiers. Figure 10 shows the Receiver Operating Characteristic (ROC) plot (true positive rate vs. false positive rate) plotted for different values of the minimum quality and HMM probability thresholds. The average of all classifiers is shown in bold. Figure 11 compares the average performance of classifiers against the training data set and the testing data set. Again, the classifiers perform significantly better against the training data as expected.

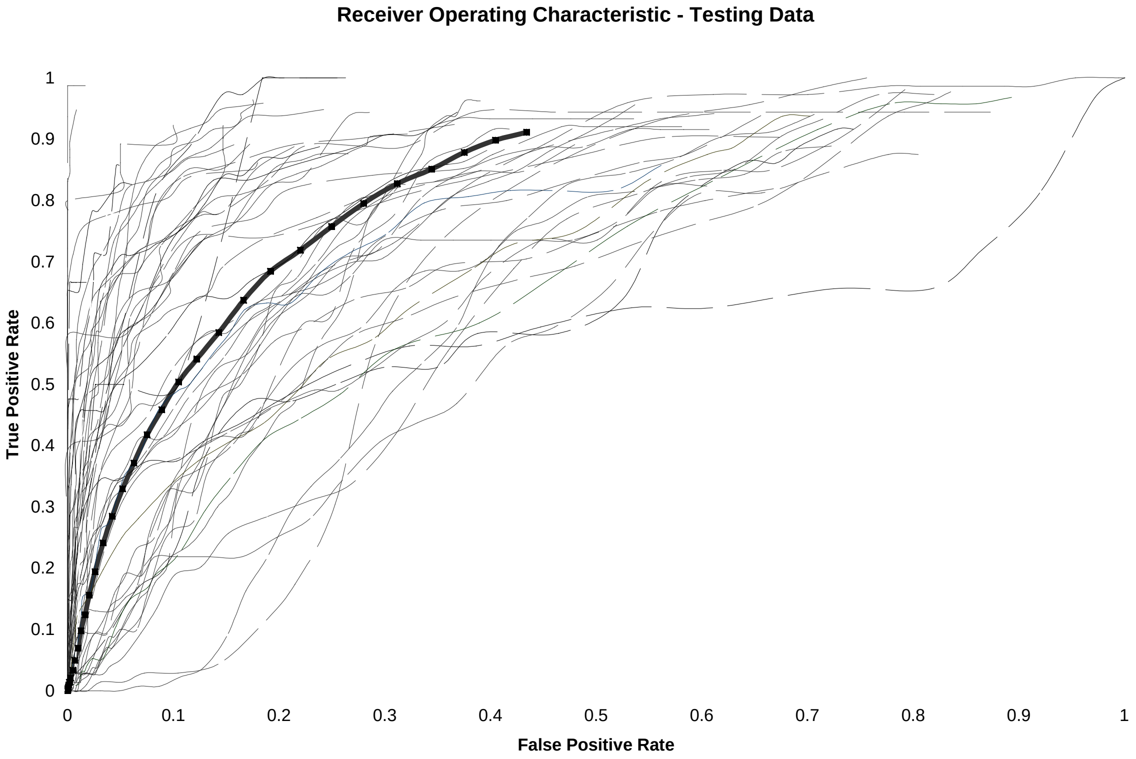

Receiver Operating Characteristic - Testing Data
True Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 10: Receiver Operating Characteristic of individual classifiers in test data

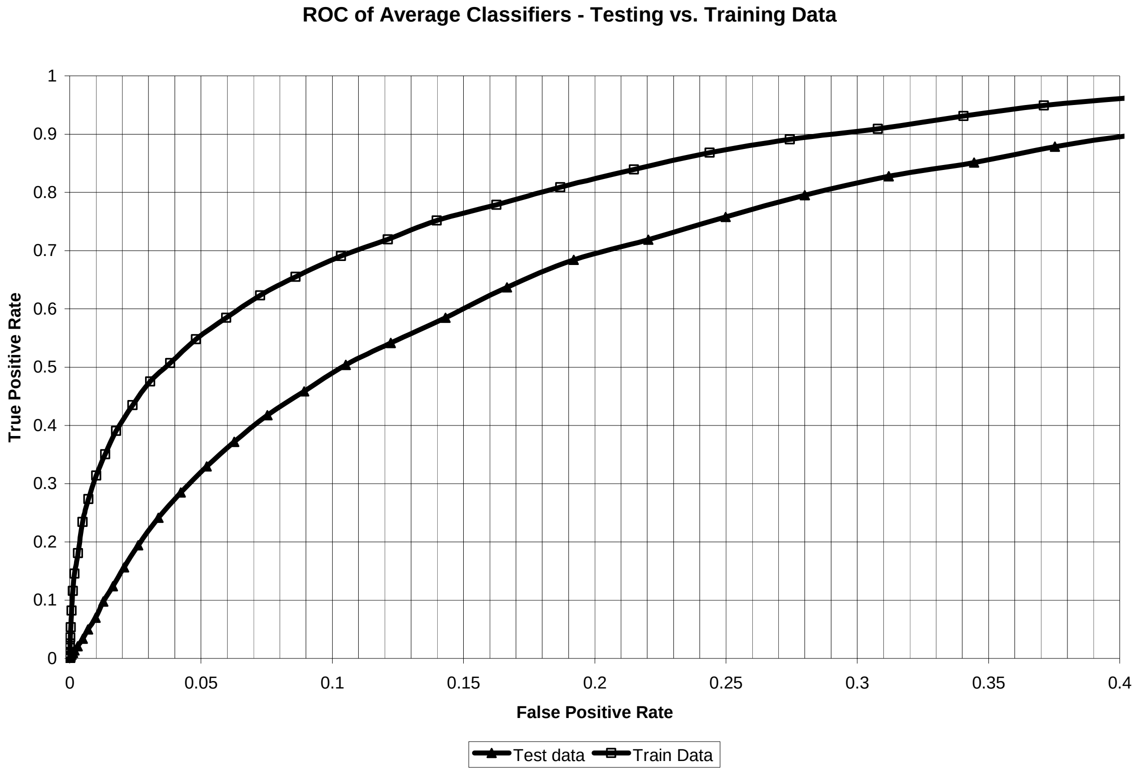


Figure 11: Receiver Operating Characteristic of Testing vs. Training Data

3.3 DISCUSSION

Overall, most of the Song Scope classifiers enjoy respectable detection rates and low false positive rates across a large matrix of vocalization types. Some classifiers performed extremely well, such as those for the Field Sparrow and the Black-and-White Warbler. In the case of the Field Sparrow, the classifier was able to detect 82% of the 78 unobserved vocalizations, with detections on all 6 test recordings, and no false positives. In the case of the Black-and-White Warbler, the classifier was able to detect 86% of the 91 unobserved vocalizations, with detections on all 5 test recordings, and no false positives.

However, some of the classifiers failed to detect previously unobserved vocalizations at all. In these cases, other classifiers scored, resulting in false positives. We believe this is due to insufficient training data where test recordings differ significantly from training recordings resulting in poor matches. To illustrate this point, we look at two examples.

First, consider the Carolina Wren classifier which failed to detect any of the 64 vocalizations on 6 test recordings. Instead, the Purple Finch classifier detected 8 of these vocalizations and the Rose-breasted Grosbeak classifier detected 7 more.

Figure 12 shows several related spectrograms. Spectrogram #1 shows a Carolina Wren vocalization present in the test data that was misclassified as a Purple Finch. Spectrogram #2 shows another Carolina Wren vocalization that was misclassified as a Rose-breasted Grosbeak. Spectrogram #3 shows a Purple Finch vocalization present in the Purple Finch classifier training data and #4 shows a

Rose-breasted Grosbeak vocalization present in the Rose-breasted Grosbeak classifier training data. Spectrograms #5 - #9 show Carolina Wren vocalizations present in the training data used to build the Carolina Wren classifier.

Notice the range of variability present in all of these Carolina Wren recordings and the fact that the two test vocalizations are not that similar to any of the 5 training vocalizations. It is therefore not surprising that the Carolina Wren depicted by spectrograms #1 and #2 failed to be detected. And while the two test recordings of Carolina Wren also do not closely resemble the Purple Finch or the Rose-breasted Grosbeak, it is not a stretch to say that they match more closely than any of the Carolina Wren training recordings.

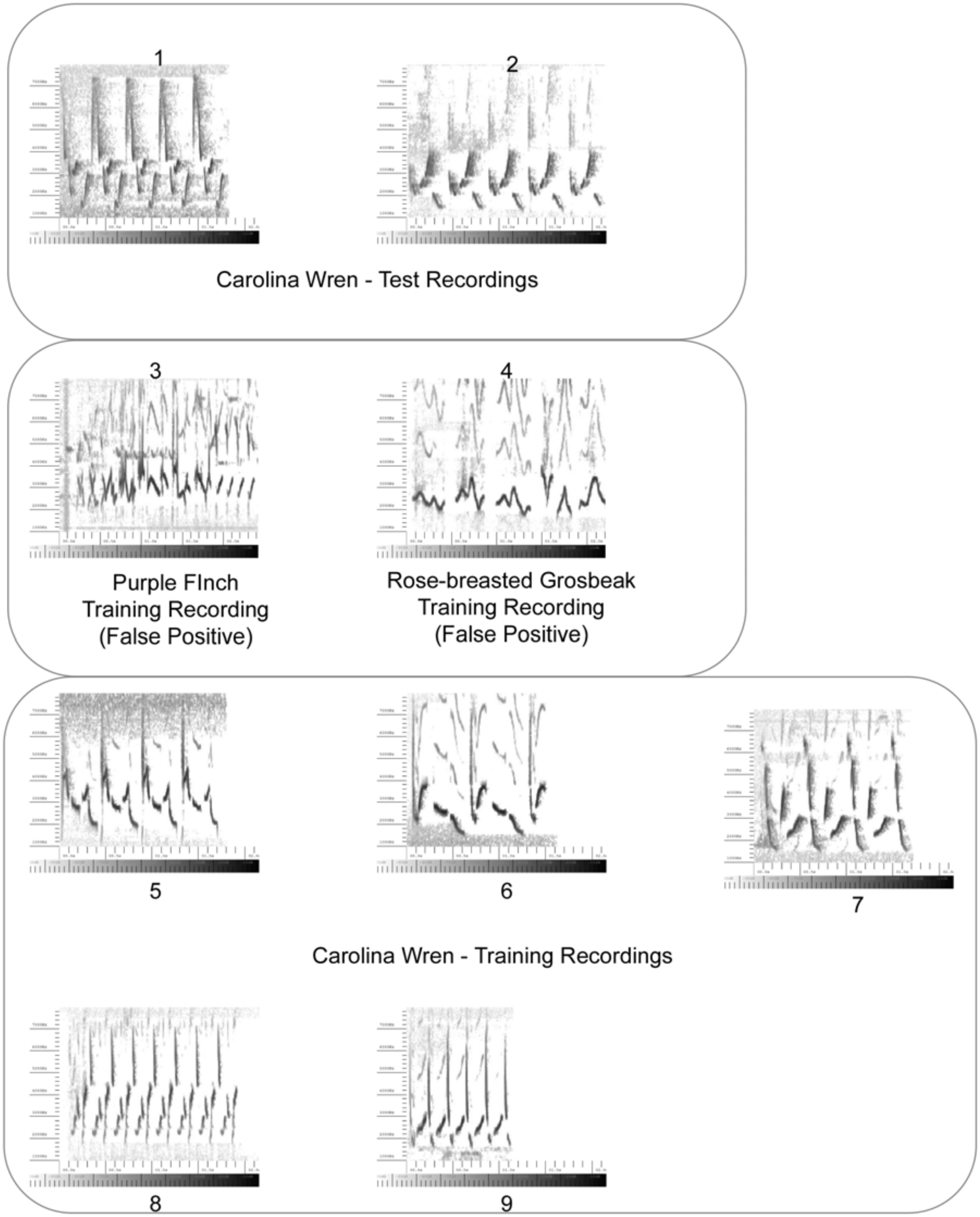


Figure 12: Spectrograms Illustrating Carolina Wren Classifier Results. The spectrograms span approximately 2.2 s and the frequency range 1.0 – 8.0 kHz.

Now consider the ROC plot for Carolina Wren, Purple Finch and Rose-breasted Grosbeak as shown in Figure 13. The Carolina Wren classifier by itself is actually above average. By itself, the classifier could detect 50% of previously unobserved Carolina Wren vocalizations with a false positive rate of only 5% across the other 53 species vocalization recordings. However, scores were relatively low compared to training data and the Rose-breasted Grosbeak classifier (more prone to false positives) scored higher.

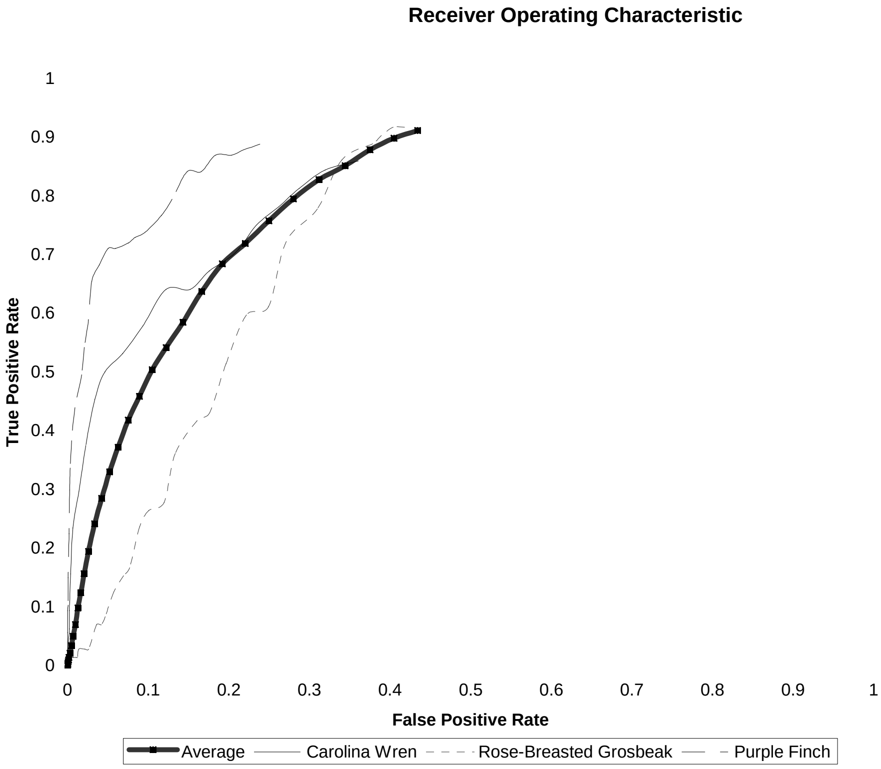


Figure 13: ROC plot of Carolina Wren compared with common false positives

Next, consider the Pine Warbler classifier which also failed to detect any of the 53 vocalizations on 5 test recordings. Instead, the Dark-eyed Junco classifier detected 6 of these vocalizations.

Figure 14 shows several related spectrograms. Spectrogram #1 shows a Pine Warbler vocalization present in the test data that was misclassified as a Dark-eyed Junco. Spectrogram #2 shows a Dark-eyed Junco vocalization present in the Dark-eyed Junco classifier training recordings. Spectrograms #3 - #8 show Pine Warbler vocalizations present in the training data used to build the Pine Warbler classifier. Notice again the range of variability present in all of these Pine Warbler recordings and the fact that the test vocalization is not that similar to any of the 6 training vocalizations. It is therefore not surprising that the Pine Warbler depicted by spectrogram #1 failed to be detected. But notice the similarity of the Dark-eyed Junco syllables in #2 and the Pine Warbler syllables in #1.

Also notice that the presence of a Pine Warbler vocalization caused the Dark-eyed Junco classifier to generate a false positive result. However, no other species (except the Dark-eyed Junco, of course) triggered the Dark-eyed Junco classifier. The false positive rate of the Dark-eyed Junco classifier will depend on the presence of Pine Warblers (and possibly other sound sources not included in the matrix) at a given site.

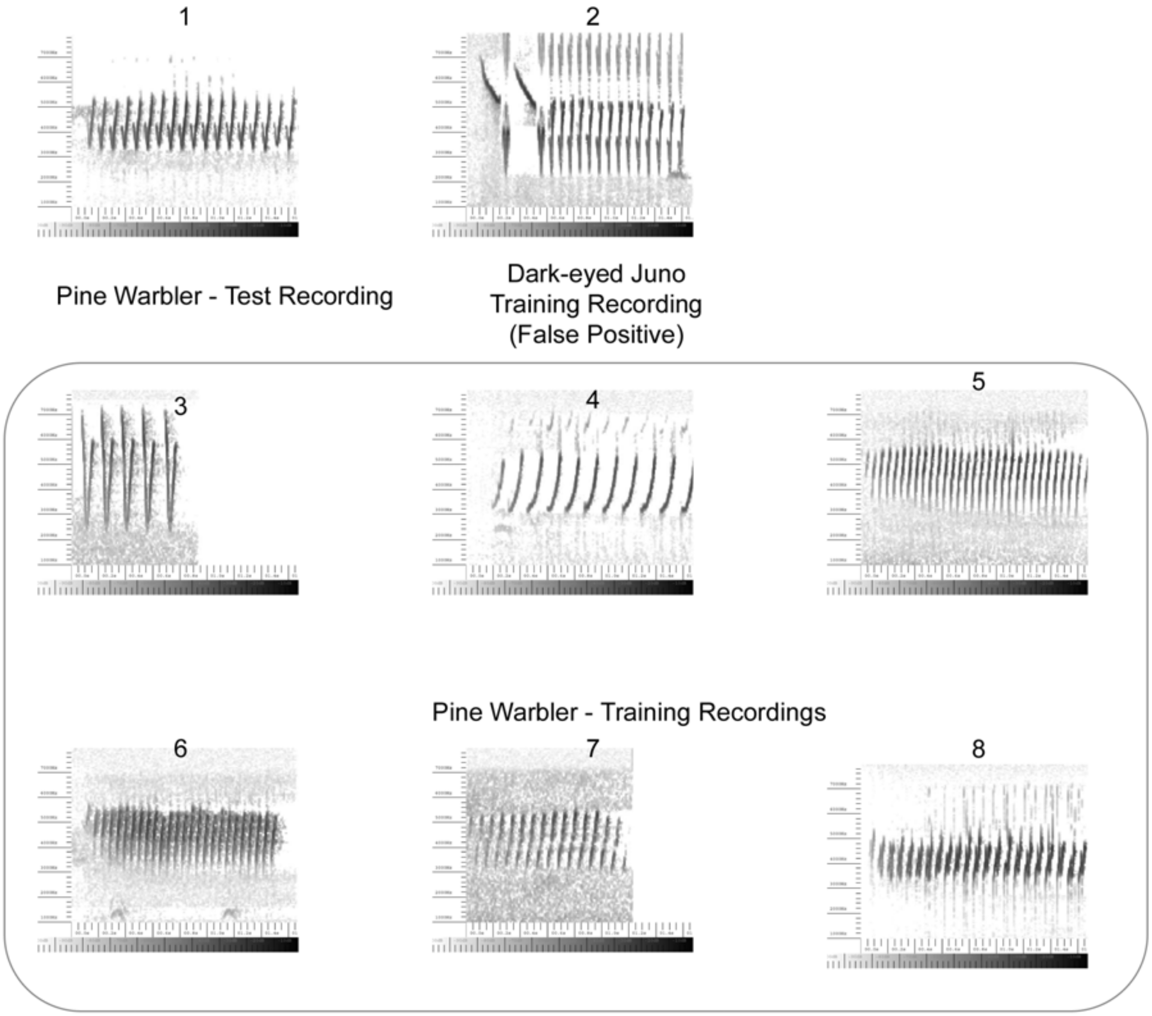


Figure 14: Spectrograms Illustrating Pine Warbler Classifier Results. The spectrograms span approximately 1.6 s and the frequency range 1.0 – 8.0 kHz.

Now consider the ROC plot for Pine Warbler and Dark-eyed Junco as shown in Figure 15. It turns out that the Pine Warbler is an exceptionally good classifier on its own capable of detecting 65% of previously unobserved Pine Warbler vocalizations with no false positives. However, the Dark-eyed Junco classifier was relatively poor prone to false positives. When considering only the highest scoring classifier, the Dark-eyed Junco classifier weakness was masked somewhat by the ability of other classifiers to score higher for their own species. But given the similarity of Pine Warbler vocalizations, the Dark-eyed Junco classifier effectively stole away the good results from the Pine Warbler classifier.

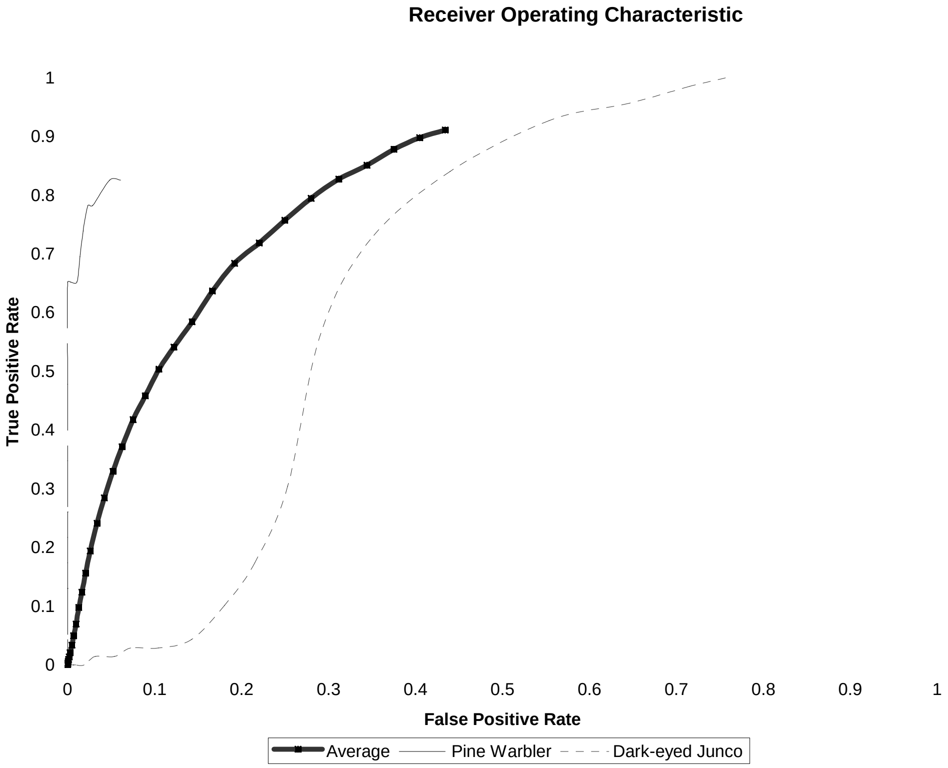


Figure 15: ROC plot of Pine Warbler vs. Dark-eyed Jucno

In the case of both the Carolina Wren and the Pine Warbler, we suspect that the range of individual variations present in these two species is not adequately reflected in the limited number of examples available in the training data. Indeed, these classifiers were built with only 5 or 6 individuals, orders of magnitude less than those used to train HMMs in speech recognition applications. While individual classifiers may have performed reasonably well, they still scored relatively low against the previously unobserved test data compared to the training data, and were out-scored by other models with similar vocalizations.

4 CONCLUSIONS

The Song Scope classification algorithms for species identification can be used to build accurate classifiers for a large range of vocalizations with limited training data. Even with only half a dozen training recordings, about half of the classifiers are capable of detecting more than 50% of previously unobserved vocalizations with false positive rates of less than 5%. A matrix of classifiers to identify common false positive sources can further improve performance. The matrix of classifiers discussed in this paper was capable of detecting 74% of unobserved individuals with false positive rates of only 0.4% on average.

The performance of individual classifiers will vary. Detection rates for some vocalizations with significant individual variability may require carefully chosen training data to detect previously unobserved individuals. False positive rates will depend on the particular constellation of competing sound sources present in the field.

The testing described in this paper was limited to generally high quality recordings and not typical field recordings captured by omni-directional microphones. However, we have demonstrated how our noise reduction and dynamic range normalization techniques compensate for this. Additional testing not detailed here suggests that these techniques do indeed work as long as vocalizations can be isolated in time and frequency from competing vocalizations (e.g. during a dawn chorus) and generally require more carefully selected parameter values for band-pass filtering and signal detection. In realistic field conditions, sensitivity rates will be slightly lower as colliding vocalizations will generally go undetected.

Low false positive rates of only 0.5% can still result in a large number of false positives when processing large quantities of field recordings, depending on the constellation of competing sound sources present. Human oversight may still be required to review detections for accuracy. However, without automatic classification, human labour would be required to review 1/0.5% = 200 times as many events. This is generally the same conclusion we arrived at in our previous publication [7] in which we demonstrated a reduction of labour in analyzing 250 hours of field recordings to 1 hour of labour to review results of automatic classification using Song Scope.

5 ACKNOWLEDGEMENTS

We would like to acknowledge the Macaulay Library at the Cornell Lab of Ornithology for providing recordings essential to our research efforts.

6 REFERENCES

1. L. Rabiner and B.H. Juang, Fundamentals of speech recognition, Prentice Hall, 321-386. (1993)
2. E. Ifeachor and B. Jervis, Digital signal processing, a practical approach, 2nd edition, Prentice Hall, 651-654. (2002)
3. M. Wilde and V. Menon, Bird call recognition using hidden Markov models, EECS Department, Tulane University. (2003)
4. S.E. Anderson, A.S. Dave and D. Margoliash, Automatic recognition and analysis of birdsong syllables from continuous recordings, Department of Organismal Biology and Anatomy, University of Chicago. (1995)
5. J.A. Kogan and D. Margoliash, Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study, Department of Organismal Biology and Anatomy, University of Chicago. (1997)
6. A.L. McIlraith and H.C. Card, Birdsong recognition using backpropagation and multivariate statistics, IEEE Transactions on Signal Processing, Vol. 45, No. 11. (1997)
7. I. Agranat, Automatic detection of Cerulean warblers using autonomous recording units and Song Scope bioacoustics software, Wildlife Acoustics, Inc. (2007)