Birdsong Recognition Using Backpropagation and Multivariate Statistics

Alex L. McIlraith and Howard C. Card, *Fellow, IEEE*

***Abstract*— An investigation has been made of bird species recognition using recordings of birdsong. Six species of birds native to Manitoba were chosen: song sparrows, fox sparrows, marsh wrens, sedge wrens, yellow warblers, and red-winged blackbirds. These species exhibit overlapping characteristics in terms of frequency content, song components, and length of songs. Songs from multiple individuals in each of these species were employed, with discernible recording noise such as tape hiss and, in some cases, other competing songs in the background. These songs were analyzed using backpropagation learning in two-layer perceptrons, as well as methods from multivariate statistics that included principal components and quadratic discriminant analysis. Preprocessing methods included linear predictive coding and windowed Fourier transforms. Generalization performance ranged from 82–93% correct identification, with the lower figures corresponding to smaller networks employing more preprocessing for dimensionality reduction. At the same time, the computational requirements were significantly reduced in this case.**

***Index Terms*—Adaptive signal processing, artificial neural networks, statistical pattern recognition.**

I. Introduction

BIRDS have been of interest to humans since history began. Both the social and ecological importance of birds are reflected in the laws we have instituted to protect them. In North America, migratory bird legislation makes it illegal to kill or collect migratory species without a permit [1]. In addition to ornithological research, a considerable number of people practice bird-watching as a hobby. It may be argued that birds are equally important to the ecology and to enhancing the quality of life. Biologists are often called upon to predict or assess the environmental impact of human activities on plants and animals. During such assessments, the biologist may have to identify and count birds in a site. Such information can then be used to assess long-term population trends in one or more species or to compare numbers between different locations and times [2]. As many of the birds in an area may be heard and not seen, it is usually convenient to rely on their sounds as a means of identification. Birds can be identified by ear because many species produce characteristic sounds. Most, but not all, such sounds are produced within the vocal tract. Of the vocal sounds, those that can be described as song are generally produced by males in a breeding or territorial context [3]. Skill in identifying songs and other sounds is essential for assessing the densities of breeding birds.

Recently, interest has arisen in the possibility of automatic species identification of bird sounds either by computer or by some form of custom dedicated hardware. An advantage of this approach is that it would not require an expert user. Such technology could also be employed in long-term environmental monitoring [4], [5]. The task of automatic species recognition is difficult because birdsongs are variable, which becomes evident when one tries to learn to identify birds by their sounds.

In this paper, we study bird species recognition using backpropagation learning in two-layer perceptrons as well as using several well-known methods from multivariate statistics. In addition, we have employed preprocessing techniques such as linear predictive coding and windowed Fourier transforms. We are interested both in the generalization accuracy after the recognition system has been trained on a dataset of recordings and on the computational requirements of the algorithms. The latter consideration is of importance if custom hardware is to be considered for field use.

II. Recognition of Birdsong: Present Work in a Historical Context

As with human speech, bird sounds can be sensibly interpreted using a frequency–time representation such as a spectrogram. In the ornithology literature, and in standard field guides, the versions of these employed are known as sonograms. These are descendants of the classical sonograph, which consisted of a bank of bandpass filters whose thresholded outputs controlled the application of ink to a strip chart in real time [6]. Narrowband and wideband options were usually available. Inked areas on the chart indicated that energy was present at a given frequency–time coordinate. Measurements of selected temporal and spectral parameters could then be extracted manually from these results. Modern software running on personal computers facilitates similar exploration through a graphical user interface [7]. Tools designed for human speech analysis are commonly used for this purpose, with analysis methods including windowed Fourier transforms, power spectral density, and linear predictive coding. In addition, biologists have considered zero-crossing analysis, autocorrelation functions, cepstral analysis, Wigner–Ville transforms, and wavelet transforms [6].

Although methods of obtaining frequency–time representations of bird sound are relatively consistent, standardized

Manuscript received June 18, 1997. This work was supported by NSERC and a University of Manitoba Fellowship to A. L. McIlraith. The associate editor coordinating the review of this paper and approving it for publication was Prof. Jenq-Neng Hwang.

The authors are with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Man., Canada R3T 5V6 (e-mail: hcard@ee.umanitoba.ca).

Publisher Item Identifier S 1053-587X(97)08059-8.

procedures have not been employed to describe temporal and spectral features [8]. Especially when these features have been extracted manually from frequency–time plots, they have been chosen rather subjectively. In obtaining these features, an element is defined as a burst of sound that is separated from other such bursts by a perceptible pause. The pause itself is termed an interelement interval. A partial list of features employed includes the frequency having the highest energy content within an element [9], the number of elements per song, the element duration, the interelement interval, the song duration, the maximum and minimum frequency content; the frequency range [10], the duration, and slope of sections of frequency-modulated sweep [11]; and the time and frequency coordinates of frequency inflections [12].

The study of extracted feature measurements has shown that invariant or stereotyped features (those varying little among populations or species) exist within songs of a given bird [13]. Furthermore, it is suggested that such features are used by birds in recognition of their own species [14], although there is some conflicting evidence; first, it has been established that different bird species use different cues to recognize members of their own species; second, features useful for discriminating between members of one group of species may not be useful in discriminating between those of a different group; third, features that are invariant within songs from a given species may not be optimal for interspecies discrimination [14], [15]. For automatic recognition, one may either employ features known to be of general importance and use a single classifier or use features that are effective in discrimination within limited subgroups of species and use specialist classifiers for each subgroup. The features of importance in these classifiers may be learned by a training procedure. Final identification, in this latter case, requires a judicious method for combining the results from the different experts, which may also be learned [16]. In this paper, we have chosen the former approach: the one of learning the best overall features in a general classifier. The modular approach, using adaptive experts [16], is left for future work.

In identifying birds by their sounds, data can be acquired easily with a directional microphone and sound recording equipment. There are also a variety of recordings available on tape and CD-ROM. In order to expand our database, we have made many of our own field recordings of the Manitoba birds in this study, but the results presented in this paper are based on commercially available recordings. Deciding on appropriate preprocessing, temporal features, and analysis methods is not as straightforward. There are many transformations that can convert a time-domain signal into a frequency–time representation, including windowed Fourier transforms and wavelet transforms [17], [18]. Windowed Fourier analysis was employed in this research, due to its wide use and conceptual simplicity. A natural extension of Fourier analysis involves obtaining statistically valid estimates of spectral content by calculating power spectral densities [20]. Another preprocessing method we have employed, which was inspired by the source-filter model commonly applied in speech analysis, is linear predictive coding (LPC) analysis [21]. Once an effective set of features is obtained, statistical methods such as stepwise discriminant analysis were also used to further reduce dimensionality prior to classification.

Artificial neural networks and a variety of multivariate statistical methods have been used for pattern recognition problems similar to that of this paper [22]. When appropriately applied, both approaches may compete effectively for the best results on a given problem [22]–[24]. In this study, two methods were used for classification. In the first, the fast Fourier transform of LPC waveforms was calculated to extract spectral features and provide data reduction. A backpropagation network was used for classification. Temporal information was provided explicitly in the form of a song-length variable and implicitly by using relatively large window sizes. In the second method, songs were parsed into periods of sound elements and silence. Spectral qualities of each element were determined using power spectral densities. The duration and number of elements and silences provided temporal information. Frequency and time information were subjected to statistical analysis, and a subset of the variables was selected for subsequent identification using both discriminant analysis and backpropagation learning.

III. Recognition Method One

Data for the study were extracted from audio tapes and compact disks that were intended for use by people wishing to identify species by song [25]–[30]. One hundred thirty three songs were chosen so that each originated from different individuals. Recordings varied in quality, with some having discernible tape hiss or the presence of other songs in the background. There was little detectable reverberation. Six species of bird native to Manitoba were chosen. These were the song sparrow (SON), fox sparrow (FOX), marsh wren (MWR), sedge wren (SWR), yellow warbler (YLW), and red-winged blackbird (RWB) [31]. They were chosen to provide a representation of both long and short songs. Songs from these species were digitized with 8-bit resolution at a 11.025-kHz sampling rate. Automatic gain control was employed, and levels were adjusted to give maximum amplitude without clipping. This reduced variation in FFT magnitudes due to amplitude variations. Recordings were left justified, and data manipulation was performed using Hypersignal Plus software.

Preprocessing required several steps. Framing was performed using a nonoverlapping Hamming window approximately 46 ms wide with 512 samples. This width is similar to frame widths used for speech recognition, which are normally in the order of 10–30 ms [32]. Sixteen time-domain coefficients were generated for a 15th-order LPC filter for each frame. Finally, a 16-point fast Fourier transform of the whitening filter coefficients was used to produce nine unique spectral magnitudes. This procedure was repeated with a 2048 sample window for all songs. The resulting data were exported from Hypersignal, with each song being represented by a number of records, each containing nine spectral coefficients. Further processing of the data was required before it could be used for training the backpropagation network. Initial investigation revealed that the overall length of the song was an important cue in species identification. Spectral and time

variables were normalized to a mean of zero and a standard deviation of one. Variables were compressed using a logistic function with a gain of unity. Two types of dataset were created. Each had records composed of ten variables derived from either the 512-sample window (9) or from the 2048-sample window (9) and the song length (1).

In this study, backpropagation without momentum or higher order derivative information was employed as the learning model [33], [34]. Experimentation with cross validation suggested a network of ten inputs, 12 hidden nodes, and six outputs was appropriate. The learning rate was set to 0.2. Target values of 0.0 and 1.0 were changed to 0.2 and 0.8 to accelerate learning [35]. Songs were divided between test and training data sets in random order. Records within songs were not randomized at this stage. The proportion of data used for testing was 25% in all runs. To evaluate performance, ten training and ten test sets were generated for each dataset type, and the network was trained with new initial weights each time. The ten test and ten training sets were generated in the same order for both datasets in order to facilitate comparisons. Preliminary training runs of 10 000 epochs indicated that 1500 epochs was sufficient for the mean sum-of-squares error to converge.

The six output values were analyzed for each of the test set records, and errors were computed. For each song, any output activation in excess of 0.6 for one of the six species was counted for that record. This technique was designed to yield consistent and somewhat conservative results. It was relatively insensitive to cases where several outputs were activated for a given frame. To establish the identity of the bird, the species class with the largest count was considered to be the winning class. If two classes were tied for the maximum count or if no class won, classification was considered to be incorrect.

IV. Recognition Method Two

Sampled bird songs were the same as those used in method one. In this case, however, parsing of a bird sound was performed using a "leaky-integrator" algorithm. Each sample of the sound was read, and if its absolute magnitude exceeded a threshold value, the integrator was incremented by an attack constant. If the magnitude fell below the threshold, the integrator was decremented by a different decay constant. The integrator was not allowed to exceed a certain limit in either the positive or negative direction. By adjusting the attack and decay parameters, it was possible to parse songs in a manner that matched human perception of elements and silences within a song. As a result of adjustment, silences that were too short in duration to be perceptible were ignored. By adjusting the threshold, sensitivity to noise was reduced.

The parsing program determined the number of elements in a song and calculated the mean and variance of both element and silence lengths within each song. In order to remove some amplitude variation prior to spectral analysis, the parsing program also normalized the amplitude of each element to the maximum scale for the data format ($+127$, -128 for 8-bit data). To verify that this procedure did not adversely affect the song, normalized songs were played back. The only perceptible difference was an overall increase in volume. Five temporal variables were produced for each song: mean and standard deviation of element length, mean and standard deviation of silence length, and the number of elements. Information concerning the temporal ordering of elements was not retained when extracting temporal or spectral information from songs.

In order to obtain a statistically robust estimate of a song's frequency content, the power spectral density was calculated for each element. The Welch method [20] was applied using a triangular time-domain window and a 16-point FFT. Variation in element lengths forced the number of FFT's (M) to vary from element to element. The squared magnitudes for the nine spectral bands, redundant negative frequencies being excluded, were accumulated. If the last window of an element contained less than 16 samples, data from this window was discarded. Once results from all FFT's were accumulated, the totals were divided by M, and their square root was taken. The final band averages were then normalized with respect to the band containing the largest average, with bands 1–9 covering the range of 0–5.125 kHz. This was done to further reduce the effect of amplitude variations in the original signal. Means and standard deviations for each of the nine spectral bands constituted the 18 spectral variables produced from each song. The data set that resulted from preprocessing contained 23 variables in 133 records. SAS software was used for all subsequent statistical analyses [36]. As we discuss below, the number of variables was reduced from 23 to eight using these statistical methods.

Preliminary examination of the correlation structure of the data indicated complex intercorrelation of variables. This in turn suggested that a smaller set of variables was expected to contain enough information to permit discrimination. A stepwise discriminant analysis was performed. The significance criterion used to enter a variable and to retain it during the elimination phase was $\alpha = 0.15$. Given that all records belonged to one of six labeled groups, it was possible to perform a single factor analysis of variance (ANOVA) with six factor levels (the species classes) using each of the eight variables. ANOVA is much like linear regression, with the difference that independent variables must be discrete. For many model-based statistical methods, including ANOVA, it is desirable that data have a normal distribution and a variance equal in each class (homoscedasticity). If such assumptions are not met, extra care must be taken in interpreting results, especially if the goal is hypothesis testing. In this study, separate ANOVA's and Scheffe's multiple comparison test [37] were calculated for each variable. Residuals were checked for normality using the Shapiro–Wilk statistic W calculated and homoscedasticity with Bartlett's test [38] using a SAS macro [38]. It is known that application of a monotonic transformation to raw data may sometimes improve the distributional or variance properties of the data. The same procedures were applied to the log and square root of the raw data. In order to visually verify that ANOVA results were reasonable, Tukey box plots [39] were also generated for each variable.

Another way of exploring the structure of data is to use ordination methods [40]. The amount of overlap among songs

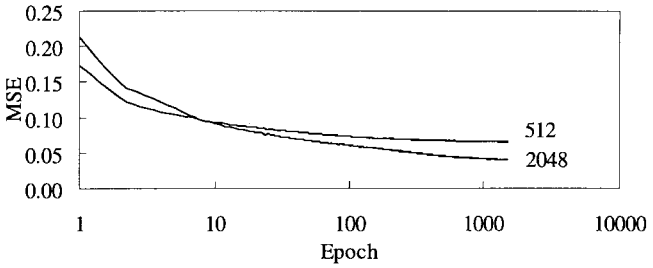


Fig. 1. Learning curves for a back-propagation network with 512 sample and 2048 sample window data sets. MSE is the mean sum-of-squares error over all training records, and averaged over 10 trials.

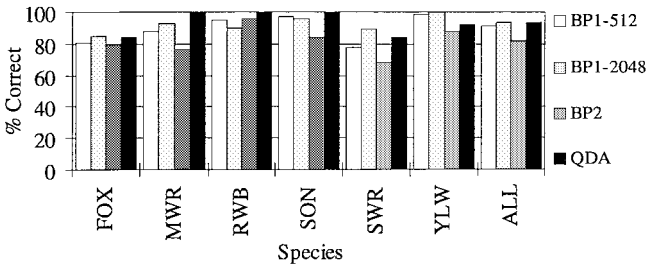


Fig. 2. Species recognition performance for three back propagation neural networks: method 1 with 512 sample (BP1-512) and 2048 sample (BP1-2048) windows, method 2 (BP2), and quadratic discriminant analysis (QDA).

of the six bird species was examined by reducing the eight variable data set to a two-dimensional (2-D) space using principal components analysis and canonical discriminant analysis. The correlation option was used for PCA so that each variable was given equal weight in the analysis regardless of its scale. CDA was then applied to the reduced data set, with prior probabilities assumed to be equal. Preliminary tests indicated that quadratic discrimination functions were necessary since covariance matrices for different classes were too dissimilar to permit pooling. A SAS program was written to divide the 133-record data set in a stratified random manner into a test set containing one record of each species and a training set that contained the remaining 127 records. Training data were used to generate discriminant functions, which were then used to classify the "unknown" observations in the test set. In order to reduce the impact of atypical records, results were based on 25 different randomizations.

In an attempt to optimize the neural network configuration and training time, training and test sets were generated in the same manner as that used in method one. The same backpropagation learning rule was used as in method one. The initial learning rate was set to 0.2, and target values in this case were set to 0.1 and 0.9. The number of inputs and outputs was eight and six, respectively. Network and training parameters were varied to find a configuration that would optimize generalization performance. To this end, the number of hidden nodes was varied from three to eight and the number of training epochs from 20–500. Classification accuracy was determined using the independent test data. Every configuration was evaluated with five trials, each utilizing different randomly selected test sets, training sets, and initial network weights. In testing, any output in excess of 0.6 was considered to be active. If any incorrect output was active or if no outputs were active, a misidentification was recorded. To make the results comparable to those derived from classification by discriminant analysis, training and test sets were generated in the same manner for both.

TABLE I
LENGTH OF SONGS; TIMES ARE IN SECONDS

Statistic	RWB	YLW	MWR	SWR	FOX	SON
Mean	0.89	1.05	1.58	1.71	2.49	2.93
Standard deviation	0.17	0.31	0.42	0.22	0.28	0.45
Number of songs	17	27	32	13	13	31

TABLE II
STEPWISE DISCRIMINANT ANALYSIS OF TEMPORAL AND SPECTRAL VARIABLES. SD IS STANDARD DEVIATION, ASCC IS AVERAGE SQUARED CANONICAL CORRELATION

Step	Variable added	Variable removed	ASCC
1	Mean element length (XE)	-	0.14
2	SD of band 4 (S4)	-	0.24
3	Mean silence length (XS)	-	0.33
4	Number of elements (NE)	-	0.42
5	SD of element length (SE)	-	0.46
6	Mean for band 3 (X3)	-	0.49
7	Mean for band 2 (X2)	-	0.52
8	Mean for band 8 (X8)	-	0.56
9	SD of band 7 (S7)	-	0.57

TABLE III
CONTRIBUTION OF VARIABLES TO THE DISCRIMINANT MODEL, INCLUDING THE PARTIAL CORRELATION COEFFICIENT FOR THE VARIABLE, AND THE CORRESPONDING F-TEST FOR SIGNIFICANCE

Variable	Partial r^2	F	Prob > F
XE	0.52	26.2	0.0001
S4	0.59	35.0	0.0001
XS	0.42	17.1	0.0001
NE	0.57	32.2	0.0001
SE	0.26	8.5	0.0001
X3	0.39	15.1	0.0001
X2	0.39	15.4	0.0001
X8	0.26	8.4	0.0001

V. RESULTS BY METHOD ONE

The mean squared error (MSE) during training was largest for the 512 sample window. This is shown in Fig. 1. Even though the differences in MSE appear to be small, they were significant over the duration of the training cycle. Two-tailed t-tests [41] comparing the mean MSE (between 512 and 2048, $n = 10$) at epoch 1500 indicate p-values $\ll$ 0.0005. This implies a negligibly small probability that the mean final MSE values were different due to chance. The network was able to recognize the six species by their song, as shown in Fig. 2. The overall performance ranged from 91–93% correct identification. Except for the SON and RWB, the mean performance was somewhat higher for the 2048 data sets than for the 512 data sets. There was considerable overlap in song length as shown in Table I. Song lengths of YLW, MWR, and SON had bimodal frequency distributions. When the raw data was examined, it was observed that certain songs were those consistently misidentified. In two cases, a song was either

TABLE IV
ANALYSIS OF VARIANCE SUMMARY. SPECIES WITH SAME LETTER IN A COLUMN HAVE MEANS THAT DO NOT DIFFER SIGNIFICANTLY (SCHEFFE'S TEST, $\alpha = 0.05$)

Species	XE	S4	XS	NE	SE	X2	X3	X8
FOX	A	A	A	A	A	A	A	A
MWR	B	A	A	A B	B	A	A B	A
RWB	B C	A B	A	C B	B	A B	C B	B
SON	B C	C B	B	C D	B	A B	C	B C
SWR	B C	C	B C	D	B	B	C	B C
YLW	C	C	C	E	B	B	C	C
Normal*	no	yes	no	no	no	no	no	yes
Homoscedastic*	no	no	no	no	no	no	no	yes

* Tests are at the $\alpha = 0.10$ level of probability.

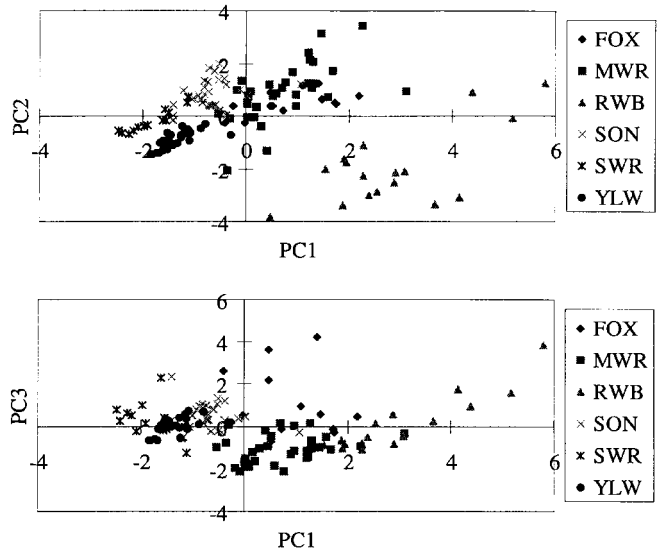


Fig. 3. First three principal components from the eight-variable data set plotted by species.

atypically long or short and was misidentified regardless of the data set in which it was included.

VI. RESULTS BY METHOD TWO

Based on the results of stepwise discriminant analysis shown in Table II, the number of variables used in subsequent analyses was reduced from 23 to eight. Using fewer variables simplified further processing. The average squared canonical correlation indicated that only small gains were likely if more than eight variables were retained, as may be observed by comparing steps eight and nine. The relative contribution of the chosen variables is indicated in Table III. Those variables with large partial correlations and F statistics contributed the most to discrimination. Note that all F statistics were significant ($p \leq 0.0001$). XE and NE were the dominant temporal variables; S4 was the dominant spectral variable. The eight variables chosen at this stage were used in all subsequent analyses.

ANOVA results and tests of assumptions are summarized in Table IV. Hypothesis tests with such data tend to be misleading [39] since assumptions of normality and homoscedasticity were violated in many cases. Although transformations were attempted, they generally improved performance on one assumption while reducing it for the other. Given that the goal at this stage was to simply explore similarities and differences among species, this was not considered to be a large problem as long as the general trends were not misleading. Inspection of the box plots confirmed that the pattern of differences among species suggested by Scheffe's test results were reasonable. Note that each variable yielded a different pattern of overlap among species, reinforcing the idea that the chosen variables each contributed different information that could be used to discriminate between species. Fox sparrows and marsh wrens, for example, could not be distinguished by the number of elements in their calls but could be separated by mean element length.

TABLE V
PCA, USING CORRELATION, FOR THE EIGHT VARIABLE DATA SET. THE PROPORTION OF STANDARDIZED VARIANCE ACCOUNTED FOR BY EACH AXIS AND THE CUMULATIVE TOTAL ARE INDICATED

Axis	Eigenvalue	Proportion	Cumulative
PC1	2.64	0.33	0.33
PC2	1.73	0.22	0.55
PC3	1.18	0.15	0.69
PC4	0.83	0.10	0.80
PC5	0.39	0.09	0.88
PC6	0.50	0.06	0.95
PC7	0.28	0.04	0.98
PC8	0.15	0.02	1.00

PCA results for the eight variables were given in Tables V and VI and Fig 3. Standardized variance accounted for by the first three components was 69%, indicating that the data had a strong underlying pattern. This also indicated that there was residual redundancy in the information provided by the original eight variables. The PCA eigenvectors provided additional useful information about the relative contribution of each variable to a given PCA axis. Based on the two vector components with the highest absolute magnitude for each axis from Table VI, PC1 appeared to be dominated by spectral variables, PC2 by temporal ones, and PC3 by both. This pattern was not particularly strong, however. Although there was considerable overlap among the six species groups, Fig. 3 indicates that red-winged blackbirds were best separated along PC1 and PC2. This group was marginally less distinct along PC3. Fox sparrows, however, were pulled away from marsh wrens along PC3. Yellow warblers, sedge wrens, and song sparrows were best separated when trends along all three axes were considered together.

Canonical discriminant analysis yielded a much stronger data structure than PCA, as was expected. This is summarized in Tables VII and VIII. Given information about the identity of

TABLE VI
EIGENVECTORS FOR THE FIRST THREE PCA AXES; THE TWO LARGEST EIGENVECTOR COMPONENTS ARE INDICATED IN BOLD TYPE

Variable	PC1	PC2	PC3
XS	-.23	0.20	**0.72**
XE	0.34	**-.50**	0.06
SE	0.37	-.27	0.13
NE	-.15	**0.62**	-.18
X2	0.36	0.23	**0.52**
X3	**0.45**	0.23	-.06
S4	0.37	0.28	-.37
X8	**-.45**	-.27	-.09

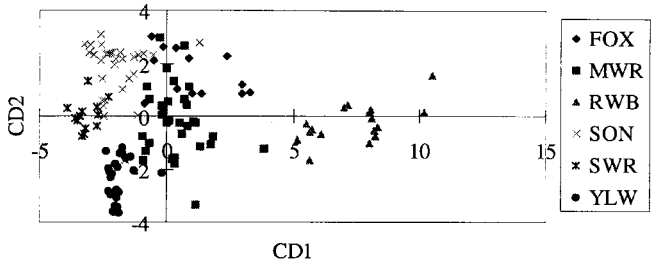


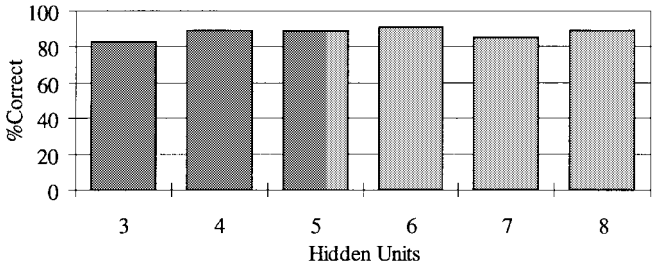
Fig. 4. First two canonical discriminants from the eight-variable data set plotted by species.

TABLE VII
CDA FOR THE EIGHT VARIABLE DATA SET. THE PROPORTION OF STANDARDIZED VARIANCE ACCOUNTED FOR BY EACH AXIS AND THE CUMULATIVE TOTAL ARE INDICATED

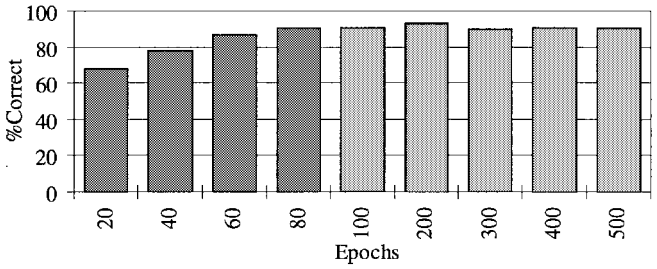
Axis	Eigenvalue	Proportion	Cumulative
CD1	10.19	0.68	0.68
CD2	2.38	0.16	0.83
CD3	1.7	0.11	0.95
CD4	0.62	0.04	0.99
CD5	0.22	0.01	1.00

TABLE VIII
EIGENVECTORS FOR THE TWO CDA AXES; THE TWO LARGEST EIGENVECTOR COMPONENTS ARE INDICATED IN BOLD TYPE

Variable	CD1	CD2
XS	-.63	0.56
XE	**0.94**	-.007
SE	**0.92**	0.12
NE	-.64	**0.58**
X2	0.73	0.48
X3	0.69	0.003
S4	0.60	0.41
X8	-.66	**-.61**



(a)



(b)

Fig. 5. (a) Effect of number of hidden units and (b) length of training period on generalization of backpropagation neural networks. Six hidden units were used in evaluation of training periods.

the classes, CDA accounted for 83% of standardized variance with only two axes and all of it with five. All variables contributed to CD1, with the time-domain variables element length and standard deviation being the most prominent. CD2, like PC3, was affected by both temporal and spectral variables. Again, CDA provided a clearer separation of species groups as shown in Fig. 4, with yellow warblers and red-winged blackbirds forming distinct groups. The other four species showed varying degrees of overlap, with fox sparrows and marsh wrens being the least distinct. Since the covariance matrices of each species group were too dissimilar to pool, quadratic discriminant functions were used in classification discriminant analysis. The results in Fig. 2 for the *test* records were excellent and showed a similar pattern of variation to those observed for method one. Overall accuracy was 93.3%. No errors were made in assigning *training* records to their proper categories.

In order to find a configuration for the backpropagation network with performance close to the optimum, a range of network configurations and training times was explored, as shown in Fig. 5, using the same eight variable data as before. Since preliminary tests indicated that training for 200 epochs resulted in reasonable generalization, networks with three to eight hidden units were evaluated. Since training was much faster than in method one, the configuration with marginally the best accuracy, having six hidden units, was chosen for further tests. The number of epochs required to achieve the best generalization on test data rose consistently until 80 epochs. After that, fluctuations were insignificant. Again, since training times in Method 2 were short, 200 epochs was chosen for maximum accuracy. Running the chosen 8–6–6 network for 200 epochs repeatedly, with one record of each species used for testing, yielded the results indicated in Fig. 2. These results were not as good as those indicated for discriminant analysis but did show similarities in trends by species. Red-winged blackbirds showed the best and sedge wrens the worst performance. The overall accuracy in this case was 82%.

VII. DISCUSSION

Songs employed in this study were recorded from individuals in a variety of locations with differing quality and amounts of background noise. Song dialects may also have been present in the data set. For example, a human listener was able to distinguish song classes among RWB songs. Since the goal was to recognize six classes of sounds from several individuals, the level of difficulty was similar to that of

speaker-independent word recognition. Given that this remains a challenging problem in human speech research, the overall performance achieved with both methods was good.

In method one, increasing the window size from 512 to 2048 samples resulted in a slight improvement in the performance of the artificial neural network. This change added some implicit temporal context beyond that provided by the song length alone. The type of information available to the network due to preprocessing is not unlike that extracted by the vertebrate auditory system. Although method one provided excellent results, considerable computation was required. Training consumed several hours on a high-performance workstation due to the dimensionality and the number of input records generated during the preprocessing. We doubt that this method will scale properly as larger sample sizes and more species are added. For these reasons, method two was investigated.

In method two, we chose to use temporal parsing of the songs followed by calculation of power spectral densities for each song element. In this way, a set of variables was generated for each song that described its temporal and spectral characteristics. Due to the variability in the number of elements within the songs, a further reduction in the number of variables describing the song was obtained by generating a number of variables that did not vary with the element count. Calculating the mean for measured parameter variables over all elements of a song summarizes the central tendency of that parameter for the song. Similarly, calculating the standard deviation for the parameter provides information about its variation within the song. In this way, a consistently small set of variables contains key frequency–time information from the song. Further dimensionality reduction employed statistical methods, in particular stepwise discriminant analysis.

Statisticians warn that care must be taken in the use of stepwise statistical procedures [42], particularly when interpreting the significance of the variables chosen by these procedures. In this study, we were less interested in the theoretical importance of variables than in their facility for discrimination. Using stepwise discriminant analysis, we were able to choose eight of the original 23 variables. Had all variables been retained, computational burdens during learning would have been increased, and performance reduced from overfitting excessive numbers of network parameters. Together, the four temporal variables selected provide information on the number of elements, their mean length, their variability, and the mean length of silent periods. The fact that their associated F statistics were large suggests that these six species can be approximately distinguished by their patterns of sound versus silence. The four spectral variables were required to sort out fine distinctions. Analysis of variance suggests that each of the eight variables on its own could be used to discriminate between subgroups of the six species. Each variable seems to contribute different information, indicating that stepwise variable selection was effective in choosing a relatively uncorrelated subset of the original variables. It is not surprising that both temporal and spectral variables were discovered to be good predictors of species identity since animals use and require both temporal and spectral information to accomplish their many tasks [14].

Many of the violations of model assumptions for ANOVA resulted from RWB songs, where only one element was detected. In addition, species containing song subgroups create bimodal or other distributions that violate the assumption of normality. Non-normality is not a serious problem as long as data are homoscedastic, and sample sizes for each class are similar [39]. Unfortunately, this was not always the case. Although the untransformed data violated ANOVA model assumptions, the patterns were not unreasonable when compared with box plots of the same data. Formal hypothesis tests were not performed. ANOVA, in the form used here, provided insight into the characteristics of the eight variables separately. PCA and CDA, however, provided a graphical two- or three-dimensional [(2-D) or (3-D)] representation from the multivariate data. Eigenvalues for both analyses suggested that there was a strong underlying structure in the data set, and that linear transformation of the eight variables into two or three retained much of the information. The two techniques differed in having temporal variables (CDA) or spectral variables (PCA) dominating the first axis. PCA and CDA plots indicate multivariate overlap not visible in the univariate ANOVA results. The ability of CDA to find linear combinations of variables that emphasized differences between species [38] permitted fewer axes to be used and increased the separation when compared with PCA. In other words, preprocessing methods were very successful in extracting relevant information for discrimination of bird species, even though important information, like the temporal order of elements within songs, was discarded. The eight variable data contained considerable multivariate overlap, indicating the recognition task was tractable but certainly not trivial.

Discriminant analysis identifies records based on characteristics of covariance matrices and works best if data are multivariate normal [36]. This means that the distributional problems that affected ANOVA results may have had a detrimental effect on discrimination. This effect is softened, however, because quadratic discrimination uses a separate covariance matrix for each species rather than a pooled estimate. The excellent results obtained indicate that distributional effects on performance were small. When the nature of the errors that discriminant analysis made in classifying were examined, five records were found to be responsible for many of the errors. Examination suggested they did indeed sound different from the others, but it was difficult to discern what the critical differences were. These samples were not incorrectly labeled but could have represented atypical or incomplete songs. They may also have represented uncommon dialects among the sample songs used in this study. Such results emphasize the need for large and representative datasets in this type of research.

The backpropagation network did not classify quite as well as quadratic discriminant analysis in method two or as well as backpropagation in method one. This, we believe, is due to suboptimal preprocessing, rather than to an undersized network for classification. Had a much larger unsupervised network been allowed to discover the best features from the underlying data, the backpropagation network may have performed better. Of course, this would have greatly increased

the overall computation and learning time as in method one. It is also possible that pathological cases were more prevalent among the test samples. In spite of the reduced performance, the smaller number of hidden units and shorter training periods in method two suggest that this approach also has its place in applications. One other description of an artificial neural network designed to recognize bird calls has been reported [43]. Backpropagation was applied to learn spectral data and appeared to obtain 87.5% accuracy using nine species. Our results are comparable with these.

VIII. Conclusions

In conclusion, the use of explicit temporal preprocessing and statistical methods to reduce data dimensionality permitted detailed exploration of species recognition from birdsong and the successful classification by both statistical and neural discriminators. Methods for backpropagation networks can trade performance for computational efficiency, enabling various applications on desktop workstations or simple custom field hardware. In future work, we will examine datasets with larger numbers of species and samples per species. We anticipate that as the recognition task becomes more difficult, the relative contribution of the parameters to the discrimination model will change, and overall performance will decline. In this case, a hierarchical network of adaptive experts [16] is preferred to initially separate different species groups. We also anticipate improved versions of these networks, perhaps with preprocessing based on unsupervised neural learning methods such as competitive learning and nonlinear principal components [35] or networks combined with hidden Markov models, to enable identification of individual birds in the wild from their sounds. This could enable the development of an inexpensive aid for studying bird populations.

Acknowledgment

The authors gratefully acknowledge insightful discussions with G. Hinton, B. Frey, D. McNeill, A. Brown, I. Khan, R. Berger, and S. Cosens on various topics related to this paper.

References

[1] Canadian Wildlife Service, Migratory Birds Convention Act, R.S., 1970, c. M-12, and the migratory birds regulations established by C.R.C., c. 1035 and amendments, Dept. Environment, Ottawa, Ont., Canada, Minister Supply Services, 1980.
[2] C. J. Ralph, G. R. Geupel, P. Pyle, T. E. Martin, and D. F. DeSante, "Handbook of field methods for monitoring landbirds," Tech. Rep. PSW-GTR-144, Pacific Southwest Res. Station, Forest Service, U.S. Dept. Agriculture, Albany, CA, pp. 1–41, 1993.
[3] D. A. Spector, "Definition in biology: The case of bird song," *J. Theor. Biol.*, vol. 168, pp. 373–381, 1994.
[4] G. McKenna, "Ecological survey at Ontario hydro's nanticoke generating station: A biodiversity conservation initiative," in *Proc. 22nd Annu. Edison Electric Inst. Biologist's Task Force Workshop*, Dallas, TX, Sept. 20–22, 1995.
[5] M. Mittelstaedt, "Ontario Hydro jolting employees into more productivity," *Toronto Globe and Mail*, p. A4, Jan. 10, 1994.
[6] T. Aubin, "Some features of time-frequency analysis and representation of animal vocalizations," in *Conf. Rep. XII Symp. Int. Bioacoust. Council*, 1992, vol. 4, pp. 59–60.
[7] P. K. McGregor, "LSI's speech workstation: A sound analysis package for IBM PC's," *Bioacoustics*, vol. 3, pp. 223–234, 1991.
[8] N. S. Thompson, K. LeDoux, and K. Moody, "A system for describing bird song units," *Bioacoustics*, vol. 5, pp. 267–279, 1994.
[9] D. M. Weary, R. G. Weisman, R. E. Lemon, T. Chin, and J. Mongrain, "Use of relative frequency of notes by Veeries in song recognition and production," *Auk*, vol. 108, pp. 977–981, 1991.
[10] T. Rich, "Microgeographic variation in the song of the sage sparrow," *Condor*, vol. 83, pp. 113–119, 1981.
[11] E. M. Date, R. E. Lemon, D. M. Weary, and A. K. Richter, "Species identity by birdsong: Discrete or additive information?," *Anim. Behav.*, vol. 41, pp. 111–120, 1991.
[12] R. B. Payne and P. Budde, "Song differences and map distances in a population of Acadian Flycatchers," *Wilson Bull.*, vol. 91, pp. 29–41, 1979.
[13] D. G. Smith, F. A. Reid, and C. B. Breen, "Stereotypy of some parameters of Red-winged Blackbird song," *Condor*, vol. 82, pp. 259–266, 1980.
[14] P. H. Becker, "The coding of species-specific characteristics in bird sounds," in *Acoustic Communications in Birds, Vol. I*, D. E. Kroodsma, E. H. Miller, and K. Ouellet, Eds. New York: Academic, 1982, pp. 213–252.
[15] D. A. Nelson, "The importance of invariant and distinctive features in species recognition of bird song," *Condor*, vol. 91, pp. 120–130, 1989.
[16] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, pp. 79–87, 1991.
[17] P. Kraniauskas, "A plain man's guide to the FFT," *IEEE Signal Processing Mag.*, vol. 11, pp. 24–35, 1994.
[18] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Mag.*, vol. 8, pp. 14–38, 1991.
[19] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE Signal Processing Mag.*, vol. 9, pp. 21–67, 1992.
[20] W. H. Press, W. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. New York: Cambridge Univ. Press, 1992.
[21] B. S. Atal, "Linear predictive coding of speech," in *Computer Speech Processing*, F. Fallside and W. A. Woods, Eds. London, U.K.: Prentice-Hall, 1985, pp. 81–124.
[22] L. Breiman, Comment, added to "Neural networks: A review from a statistical perspective," in *Statist. Sci.*, B. Cheng and D. M. Titterington, Eds., vol. 9, pp. 2–54, 1994.
[23] B. D. Ripley, "Neural networks and related methods for classification," *J. R. Stat. Soc. B.*, vol. 56, pp. 409–456, 1994.
[24] B. Cheng and D. M. Titterington, "Neural networks: A review from a statistical perspective," *Statist. Sci.*, vol. 9, pp. 2–54, 1994.
[25] M. Brigham, *Bird Sounds of Canada.* Mount Albert, Ont., Canada: Holborne.
[26] D. J. Borror, *Songs of Eastern Birds.* New York: Dover, 1970.
[27] _____, *Common Bird Songs.* New York: Dover, 1967.
[28] L. Elliot and T. Mack, *Wild Sounds of the Northwoods.* Ithaca, NY: NatureSound Studio, 1990.
[29] R. K. Walton and R. W. Lawson, *Birding by Ear (Eastern/Central)—A Guide to Bird-Song Identification.* Boston, MA: Houghton-Mifflin, 1989.
[30] P. P. Kellogg, R. T. Peterson, and W. W. H. Gunn, *A Field Guide to Western Bird Songs.* Boston, MA: Houghton-Mifflin, 1975.
[31] M. B. Robbins, M. J. Braun, and E. A. Tobey, "Morphological and vocal variation across a contact zone between the Chickadees *Parus Atricapillus* and *P. carolinensis*," *Auk*, vol. 103, pp. 655–666, 1986.
[32] R. P. Lippmann, "Review of neural networks for speech recognition," *Neural Comput.*, vol. 1, pp. 1–38, 1989.
[33] D. E. Rumelhart, F. E. Hinton, and R. J. Williams, "Learning representations by back-propagation of errors," *Nature*, vol. 323, pp. 533–536, 1986.
[34] J. L. McLelland and D. E. Rumelhart, *Explorations in Parallel Distributed Processing.* Cambridge, MA: MIT Press, 1989.
[35] S. Haykin, *Neural Networks: A Comprehensive Foundation.* New York: Macmillan, 1994.
[36] *SAS User's Guide: Statistics Version 5 ed.*. Cary, NC: SAS Inst., 1985.
[37] J. Neter and W. Wasserman, *Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Designs.* Homewood, IL: R. D. Irwin, 1974.
[38] M. Friendly, "Bartlett's test for homogeneity of variances: A SAS macro," Dept. Psychol., York Univ., Toronto, Ont., Canada, 1995; http://www.math.yorku.ca/SCS/friendly.html.
[39] J. W. Tukey, *Exploratory Data Analysis.* Reading, MA: Addison-Wesley, 1977.
[40] E. C. Pielou, *The Interpretation of Ecological Data: A Primer on Classification and Ordination.* New York: Wiley, 1984.
[41] G. K. Bhattacharyya and R. A. Johnson, *Statistical Concepts and Methods.* New York: Wiley, 1977.
[42] F. C. James and C. E. McCulloch, "Multivariate analysis in ecology and

systematics: A panacea or Pandora's box?," *Annu. Rev. Ecol. Syst.*, vol. 21, pp. 129–166, 1990.

[43] T. Ashiya and M. Nakagawa, "A proposal for a recognition system for the species of birds receiving birdcalls—An application of recognition systems for environmental sound," *IEICE Trans. Fundamentals*, vol. E76-A, pp. 1858–1860, 1993.

Alex L. McIlraith received the B.Sc. degree in ecology and M.Sc. degrees in both botany and electrical engineering from the University of Manitoba, Winnipeg, Man., Canada.

He is a research associate with the University of Manitoba and coordinates hardware design activities at a local company in Winnipeg. His current research interests include signal processing and applications of electronic technologies to biological problems. Related activities include the recording and analysis of animal sounds.

Howard C. Card (F'94) received the Ph.D. degree from the University of Manchester, Manchester, U.K., in 1971.

He has held academic appointments at Manchester; the University of Manitoba, Winnipeg, Man., Canada; the University of Waterloo, Ont., Canada; and Columbia University, New York, NY, and has spent a sabbatical year at Oxford University, Oxford, U.K. He has worked on device physics and modeling, parallel VLSI computations, and artificial neural networks. His current research interests are in microelectronic and software systems design employing artificial neural learning and its applications in adaptive web agents and user interfaces, adaptive consumer appliances, mobile robots, and educational toys. He is also interested more generally in the relationship between biology, physics, and computation.

Dr. Card has received several teaching awards including the Stanton Award for Excellence in Teaching and the UMFA-UTS Award. He also holds several research awards including the Rh. Institute Award for Multidisciplinary Research, the Sigma Xi Senior Scientist Award, the NSERC E.W.R. Steacie Memorial Fellowship, and the ITAC-NSERC Award for Academic Excellence.