Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models

Vlad M. Trifa,[a] Alexander N. G. Kirschel,[b] and Charles E. Taylor[c]
Department of Ecology and Evolutionary Biology, University of California Los Angeles, 621 Charles Young Drive South, Los Angeles, California, 90095

Edgar E. Vallejo[d]
Department of Computer Science, ITESM-CEM, Carretera Lago de Guadalupe km. 3.5, Col. Margarita Maza de Juárez, Atizapán de Zaragoza, 52926, Estado de México, México

(Received 21 February 2007; revised 8 January 2008; accepted 9 January 2008)

Behavioral and ecological studies would benefit from the ability to automatically identify species from acoustic recordings. The work presented in this article explores the ability of hidden Markov models to distinguish songs from five species of antbirds that share the same territory in a rainforest environment in Mexico. When only clean recordings were used, species recognition was nearly perfect, 99.5%. With noisy recordings, performance was lower but generally exceeding 90%. Besides the quality of the recordings, performance has been found to be heavily influenced by a multitude of factors, such as the size of the training set, the feature extraction method used, and number of states in the Markov model. In general, training with noisier data also improved recognition in test recordings, because of an increased ability to generalize. Considerations for improving performance, including beamforming with sensor arrays and design of preprocessing methods particularly suited for bird songs, are discussed. Combining sensor network technology with effective event detection and species identification algorithms will enable observation of species interactions at a spatial and temporal resolution that is simply impossible with current tools. Analysis of animal behavior through real-time tracking of individuals and recording of large amounts of data with embedded devices in remote locations is thus a realistic goal.
 [DOI: 10.1121/1.2839017]

PACS number(s): 43.80.Ka, 43.72.Ne, 43.60.Lq [JAS] Pages: 2424–2431

I. INTRODUCTION

Bird songs play an important role in species-specific communication, primarily for mate attraction and territory defense (Catchpole and Slater, 1995). Species identification based on acoustic communication is particularly important in rainforest environments because of the limited visibility caused by dense vegetation. Within such environments there is a need to efficiently distinguish between species' songs in order to understand how species interact. While there is no substitute for the experienced human observer, automated monitoring, using sensor networks, is increasingly recognized for its unobtrusiveness, constant alertness, and extended presence for localization and situational awareness (Sczewczyk *et al.*, 2004). The required software methods to automatically analyze bird vocalization from using *embedded networked sensors* have only very recently become available. This study focuses on a class of methods called *hidden Markov models* to automate recognition of birds in a Mexican rainforest.

Species recognition based on objective analysis of vocalization has been applied to a wide variety of animals, including whales (Scheifele *et al.*, 2005), marmots (Blumstein and Munos, 2005), grasshoppers (Chesmore, 2004), and birds (Fagerlund, 2004). Previous studies have used a variety of methods for species recognition. At the simplest level, that of filtering for species or individual identification, several approaches have been explored. These approaches include: multivariate statistical analysis of song properties (Clark *et al.*, 1987; McIlraith and Card, 1997), cross correlations of spectra (Clark *et al.*, 1987), artificial neural networks with or without backpropagation (Ashiya and Nakagawa, 1993; McIlraith and Card, 1997), dynamic time warping (Anderson *et al.*, 1996; Kogan and Margoliash, 1998), Kohonen self-organizing maps (Somervuo, 2000), and sinusoidal modeling of syllables (Härmä, 2003).

Some researchers have attempted to describe and separate species' song according to temporal and spectral features of the acoustic signal. Typically, from the large number of features that can be extracted from bird songs, a small percentage is generally sufficient for correct classification. However, these distinguishing features are not invariant, but rather change from population to population, based on community structure (Nelson, 1989). Other variants of feature analysis include metrics to quantify song similarity (Tchernichovski *et al.*, 2000), linear discriminant functions (Fager-

[a]Current address: Institute for Pervasive Computing, ETH Zurich, Haldeneggsteig 4, 8092 Zurich, Switzerland. Electronic mail: vlad.trifa@ieee.org
[b]Electronic mail: kirschel@ucla.edu
[c]Electronic mail: taylor@biology.ucla.edu
[d]Electronic mail: vallejo@itesm.mx

lund and Härmä, 2004), and data mining (Vilches *et al.*, 2006) to identify rules for maximum discriminative power.

In the last decade, most of these efforts have benefited from progress in the field of automated speech recognition, in particular with hidden Markov models (HMMs) (Kogan and Margoliash, 1998; Kwan *et al.*, 2004; Wilde and Menon, 2003). These methods are usually more efficient than knowledge-based recognition methods, especially when used in noisy environments. This method is commonly used in automated speech recognition and is generally known to be effective for modeling time-series data (Rabiner, 1989). Vilches *et al.* (2006) presents a comparison between HMMs and Bayes classifiers where the features used had been identified by data mining for the ability to discriminate among three species of antbirds. They found that both methods performed very well, over 90% of correct classification. Data mining is a more statistically optimal method for classification, but the drawback is that it requires extensive data preparation. Most previous work used laboratory recordings, which simplifies the analysis as it increases signal-to-noise ratio (SNR), minimizing any background noise interference. In principal, HMMs have the advantage of not requiring expert human intervention for preprocessing (feature extraction) and thus can run in real-time [in Vilches *et al.* (2006), features were extracted manually], while still exhibiting comparable performance to data mining methods.

The study reported here employs HMMs to distinguish among the songs of antbirds from a tropical rainforest in Southern Mexico. The goal is to provide a standard, yet efficient, method of bird species classification that can be easily implemented, customized, and run in real-time using off-the-shelf hardware and software. In contrast with most previous work, whole songs rather than individual syllables were used in the classification. In addition, the role of several parameters of the HMM analysis was investigated, particularly the quality of data used for training and for testing, and the minimal number of samples required to train the models while ensuring sufficiently good recognition.

II. METHODS

A. Data collection and filtering

Bird songs were recorded at the Estacion Chajul in the Reserva de la Biosfera Monte Azules, in Chiapas, Mexico (approximately 16°6′44″ N and 90°56′27″ W), during June 2005 and February 2006. Recordings were made using a Sennheiser ME67/K6 directional microphone and a Marantz PMD670 digital recorder directly onto Compact Flash cards at a sample rate of 44.1 kHz and a resolution of 16 bits per sample. Recordings were collected at various distances from subjects, resulting in a high variation in SNR to replicate conditions of recordings that would be collected from remote recording devices. Full recordings were processed with a high-pass filter to remove noise in the low frequencies with the filtering function implemented within RAVEN (high-pass filter with cutoff frequency 400 Hz). Afterward, single songs were manually extracted and classified in order to provide the set of songs used for the training of the models (typical length of each song: between 2 and 3.6 s, depending on the species). Each song was initially assigned a value (A–E) to reflect the quality of the recording based on subjective estimation of its SNR, where highest quality songs were assigned an A and the poorest quality ones were assigned an E. Songs rated A, B, or C are considered as high quality, while D or E are considered too noisy. Training samples (TS) refers to the number of random samples (grades A–E) used to train the network; clean training samples (CTS) refers to only clean recordings (grades A–C) used for training. Every individual antbird song was extracted from the recordings using the RAVEN 1.3 program (Charif *et al.*, 2006), and then used for the HMM analysis as a training or test song. Such a subjective measure of the recording quality was chosen—as opposed to a quantitative SNR estimated (e.g., ratio of peak power versus noise power)—because there is interest to see how such a system could perform mainly with songs that human experimenters can discern, while discarding the ones that would be too weak to be recognized by human observers (as it is difficult to obtain any meta information about the caller, e.g., the location, species, and other individual characteristics). The real-time detection and extraction of full songs in noisy environments is not addressed in this article, therefore interested readers are invited to consult Ali *et al.* (2007) for details about how this procedure would be implemented.

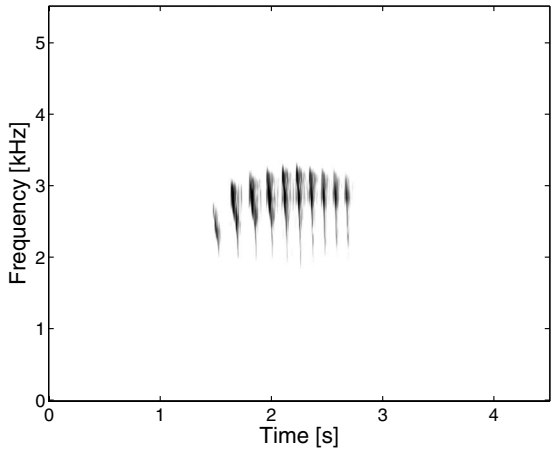

FIG. 1. Spectrograms of songs of species used in this study: Dusky Antbird (DAB) (*Cercomacra tyrannina*).

Five antbird species were included in the analysis. These are the typical antbirds (Thamnophilidae): Great Antshrike (GAS) (*Taraba major*), Barred Antshrike (BAS) (*Thamnophilus doliatus*), Dusky Antbird (DAB) (*Cercomacra tyrannina*), Dot-winged Antwren (DWAW) (*Microrhopias quixensis*), and the ground antbird (Formicariidae) Mexican Antthrush (MAT) (*Formicarius analis*). Spectrograms of the songs of these species are illustrated in Figs. 1–5.

B. Hidden Markov models

A HMM is a statistical tool that can model a discrete-time dynamical system described by a Markov process with unknown parameters. In the context discussed here a bird song can be considered as a sequence of observations produced by such a dynamical system. A simple representation

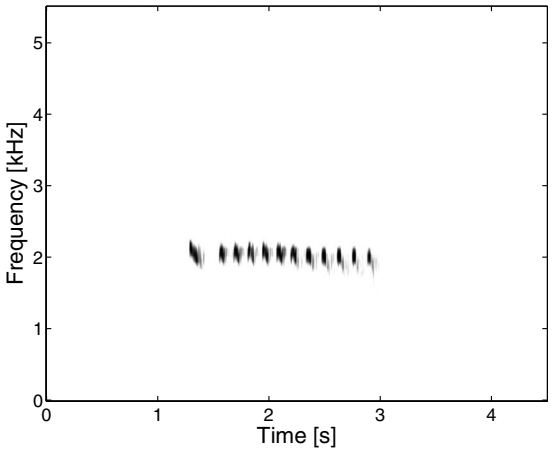

FIG. 2. Spectrogram of song from Mexican Antthrush (MAT) (*Formicarius analis*).

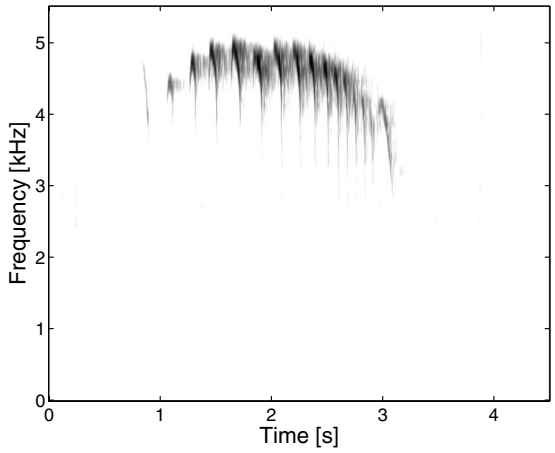

FIG. 4. Spectrogram of song from Dot-winged Antwren (DWAW) (*Microrhopias quixensis*).

where each state of this Markov process corresponds to an observable event is too restrictive to be applied to real world problems, because of the difficulty in relating the states with observable events. Accordingly, the model is extended to allow each observation to be modeled as a probabilistic function of the state. In this paradigm, the sequence of observations will not tell anything about the state sequence that generated them, because the sequence of states that produced the observation is not visible. For each bird species, a different HMM is used to model the temporal progression of the acoustic features of its species, and recognition can be done by inferring from observed data which HMM is the most likely one to produce the given sequence of observations. The challenge becomes to estimate the parameters of the HMM from sample observations (*training observation*), and then use the estimated parameters to infer the probability that a dynamical system produced a given observed sequence, so as to find the most probable sequence of unobservable states that produced the observed sequence (this sequence can be used as a classification method).

Opposed to pattern matching, the concept of knowledge-based methods requires one to analyze the signal in order to quantify a set of features at regularly spaced discrete time steps, and then compare songs based upon this information. For example, a vector of features is extracted from a signal at discrete time step (each of these vectors will be called an *observation*), and these vectors will result in a time series. Afterwards, an HMM for each class to recognize (here each class refers to a different bird species) is used to model the temporal evolution of the features of its class, and recognition is done by looking at which HMM is the most likely to produce a given sequence of observations. The ability to model such time series with subtle temporal structure makes HMMs an interesting alternative to pattern-matching methods, as HMMs are very efficient for modeling patterns of varying length.

Experimental evidence has shown that the human ear does not resolve frequencies linearly across the audio spectrum, and empirical evidence suggests that designing a front-end to operate in a similar manner typically yields better recognition performance than a simple Fourier transform. The principle of this approach, called *mel-frequency cepstral coefficients* (MFCC), is similar in principle to a Fourier transform, but the frequency bands in the MFCC filters are equally spaced in mel frequency, a linear-log function of frequency which explains the sensitivity of human pitch percep-

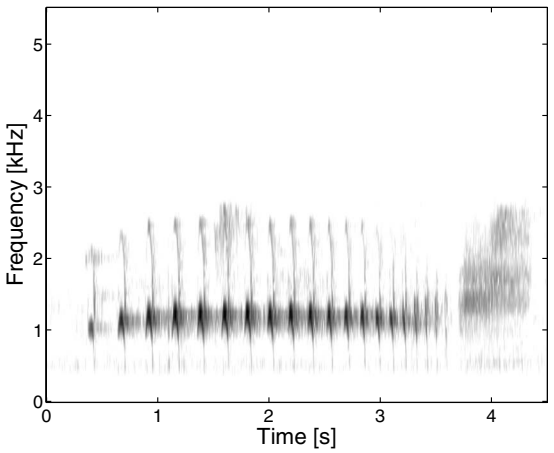

FIG. 3. Spectrogram of song from Great Antshrike (GAS) (*Taraba major*).

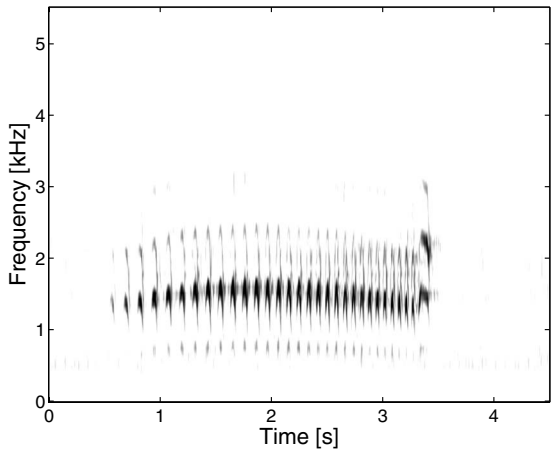

FIG. 5. Spectrogram of song from Barred Antshrike (BAS) (*Thamnophilus doliatus*).

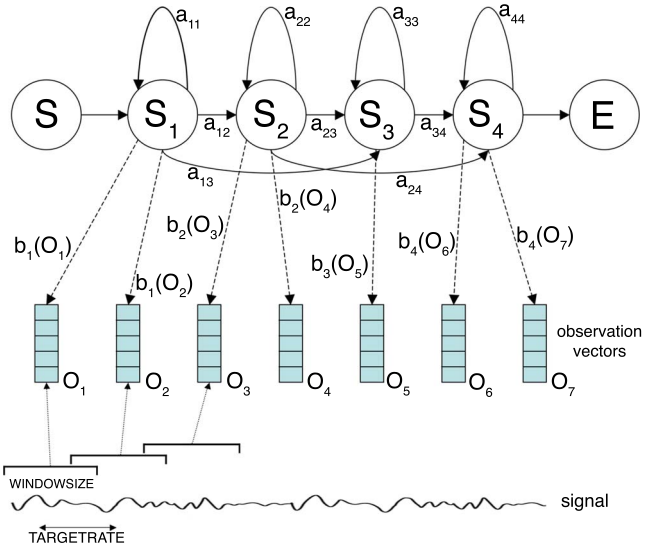

FIG. 6. (Color online) Illustration of the HMM theory. From the original acoustic signal, observation vectors (O_i) are extracted from blocks of duration WINDOWSIZE ms. The interval between two observations is TARGETRATE ms. The transition probability from state i to state j is a_{ij}, and the probability of observing a vector k while in the state i, is denoted by $b_i(O_k)$. Illustration inspired from Young *et al.* (2002).

tion. Such a filter bank can be implemented by taking the magnitude of the Fourier transformation for a window of data, and then multiplying the resulting magnitudes coefficients using triangular filters for different frequency bands. Thus, each band will contain a weighted sum representing the spectral magnitude in the corresponding particular channel.

An alternative method to generate the observation vectors from a signal is to use *linear predictive coding* (LPC), which is commonly used in communication systems. Usually, there exists a correlation between the samples in a segment of acoustic data, and the idea behind LPC is to encode an acoustic segment as a set of coefficients in a given equation that allows one to predict the value of all samples in the segment according to the preceding samples ones. The coefficients are then chosen so that they minimize the error between the actual values of the samples and the values predicted by the equation.

1. HMMs theory

Formally, a HMM is a five-tuple $(\mathbf{\Omega}_X, \mathbf{\Omega}_O, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$, where $\mathbf{\Omega}_X = [S_1 \ldots S_N]$ is a finite set of N distinct states, while $\mathbf{\Omega}_O = [v_1 \ldots v_k]$ is the set of possible observation symbols (the alphabet of observation), where $\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ denotes the parameters of the hidden Markov chain, with $\boldsymbol{A}_{N\times N}$ the transition probabilities matrix, $\boldsymbol{B}_{N\times k}$ the probabilities of observing each symbol for each state, and $\boldsymbol{\pi}_{1\times N}$ the distribution of the initial state (see illustration Fig. 6).

The state of the system can change at discrete time steps, and the next state is defined by the matrix of transition probabilities A. The state of the system at time t is denoted q_t. The Markov chain assumption states the probability of passing from state S_i into state j at time t depends only on the current state only and not on the previous state changes, that is

$$P(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \ldots) = P(q_t = S_j | q_{t-1} = S_i) = a_{ij}. \quad (1)$$

Also, this probability is considered as being independent of the time, and a_{ij} is called the state transition (from state i to state j) coefficient (A is the matrix composed of a_{ij} for all i,j), having the following properties:

$$0 \leqslant a_{ij} \leqslant 1, \quad \sum_{j=1}^{N} a_{ij} = 1. \quad (2)$$

The probability of observing a particular multidimensional configuration is modeled using Gaussian mixture models, as follows

$$b_j(O_t) = \sum_{m=1}^{M} c_{\mathrm{dim}} N(O_t; \mu_{\mathrm{dim}}, \Sigma_{\mathrm{dim}}), \quad (3)$$

where M is the number of mixtures, c_{dim} is the weight of mixture m, $N(O_t; \mu_{\mathrm{dim}}, \Sigma_{jm})$ is a Gaussian density with mean vector $\mu_{1\times P}$ and covariance matrix $\Sigma_{P\times P}$, both of same dimension as the observation vector O_t, that is P.

In practice the parameters of HMMs are unknown and have to be estimated from data during the training phase. The choices made for these parameters are determinant for performance of the model, and thus should reflect as reliably as possibly the data to be modeled.

2. The three problems of HMMs

There are three major problems to solve when using HMMs for real-world applications. The reader is invited to consult Rabiner and Juang (1993) and Rabiner (1989) for detailed explanations about mathematical description of HMMs and solutions used for each of these problems, which are presented here only for general understanding:

a. Problem 1. Given an observation sequence $\boldsymbol{O} = O_1 \ldots O_T$ and a model $\boldsymbol{\lambda} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$, how to estimate the probability $P(\boldsymbol{O}|\boldsymbol{\lambda})$ that the sequence $\boldsymbol{O}$ has been produced by the model $\boldsymbol{\lambda}$? In other terms, how to score each model according to how well it matches the observation?

A straightforward solution to this problem is simply to enumerate every possible sequence of length T and compute their probability to be observed (assuming statistical independence between observations), that is

$$P(\boldsymbol{O}|\boldsymbol{Q}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} P(O_t | q_t, \boldsymbol{\lambda}) = b_{q_1}(O_1) \ldots b_{q_T}(O_T). \quad (4)$$

The probability of the observation is simply a sum of this joint probability over all possible sequences of length T,

TABLE I. Average recognition performance (% of correctly classified songs) for five species of antbirds using the method MFCC_0_A_D, thus the size of the observation vector is 39. Numbers in parentheses are the standard deviation of the results over 50 trials. TS refers to the number of random samples that were to train the network (grades A–E); CTS refers to the number of clean recordings (grades A–C only). All results refer to tests against the remaining grade A–E songs.

Species	Total songs	10 TS	20 TS	50 TS	100 TS	10 CTS	50 CTS
Barred Antshrike (BAS)	360	95.2 (3.2)	96.23 (2.1)	97.7 (0.9)	97.8 (0.9)	95.3	98.4
Dusky Antbird (DAB)	1333	93.1 (3.4)	95.05 (1.83)	96.2 (1.0)	96.5 (1.0)	85.9	95.2
Dot-winged Antwren (DWAW)	237	96.3 (1.7)	96.3 (1.45)	96.8 (1.0)	97.1 (1.7)	99.5	99.5
Great Antshrike (GAS)	525	91.0 (3.6)	92.58 (2.22)	93.3 (2.3)	94.2 (1.7)	81.4	90.5
Mexican Antthrush (MAT)	913	92.5 (4.1)	93.99 (2.59)	96.3 (1.3)	96.8 (1.2)	91.3	95.6
Overall performance (%)		93.04 (1.65)	94.59 (1.04)	95.97 (0.63)	96.38 (0.51)	88.56	95.19

$$P(\boldsymbol{O}|\boldsymbol{\lambda}) = \sum_{\boldsymbol{Q}} P(\boldsymbol{O}|\boldsymbol{Q},\boldsymbol{\lambda}) \cdot P(\boldsymbol{Q}|\boldsymbol{\lambda}) = \sum_{q_1 \cdots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \ldots a_{q_{T-1} q_T} b_{q_T}(O_T). \quad (5)$$

However, using Eq. (5) to compute $P(\boldsymbol{O}|\boldsymbol{\lambda})$ is not feasible in practice, as the computational complexity of this procedure is far too big to be useful, so much more efficient methods to compute this value have been proposed.

b. Problem 2. Given an observation $\boldsymbol{O}$ and a model $\boldsymbol{\lambda}$, how to choose the "correct" state sequence $\boldsymbol{Q}=q_1 \ldots q_T$ that best explains the observation $\boldsymbol{O}$? Here, the problem is to attempt to discover the hidden part of the model.

c. Problem 3. How to adjust the model parameters $\boldsymbol{\lambda}$ in order to maximize $P(\boldsymbol{O}|\boldsymbol{\lambda})$? This is a crucial aspect of HMMs as it is the training procedure, i.e., adapt the parameters of $\boldsymbol{\lambda}$ to fit optimally each of the samples of the training data set.

C. Hidden Markov model toolkit (HTK)

As described in Young *et al.* (2002), HTK is a package of libraries and tools written with the C programming language that provides sophisticated facilities for designing, training, and testing of HMMs, as well as analyzing the results. HTK has been chosen in this project as it is one of the most common and flexible implementation of HMMs and has been successfully applied in many automatic speech recognition applications. Also, the source code is freely available, thus HTK could be compiled and run on customized embedded computers.

D. Model parameters

The experiments presented used two standard feature extraction methods proposed by HTK, MFCC, and LPC were used. Observations are extracted from a segment of the signal of length WINDOWSIZE, and a segment is extracted every TARGETRATE.

To improve the recognition performance, one can add to the coefficients generated using MFCC or LPC, other information extracted from the signals, and some of these terms were also investigated here. These were: the logarithm of the signal energy (_E), the first-order regression coefficients (the rate of change) of the basic static parameters (the LPC/MFCC coefficients) between two consecutive time frames, (delta coefficients _D), and second-order regression coefficients (the second time derivative, acceleration _A), and the zeroth-order cepstral coefficient (_0). The other parameters are set to HTK default values (the filterbank has 26 channels and 12 MFCC coefficients are output, LPC order is 12).

Practical use of HMMs is still an art, as much as it is a science. One of the difficulties in using HMMs is that the procedure requires parameters for which there is no algorithm to obtain an optimal value based on the data at hand. From pilot studies, the following settings were found to best represent the data: WINDOWSIZE=25 ms and TARGETRATE =15 ms. If the songs differ greatly in length across species, then increasing the state length (increase the TARGETRATE parameter, so a state is responsible for a longer segment of the song) should be considered. The resulting state transition probabilities will be more different across models, resulting in less ambiguous recognition. The number of states was set to 5, based on the simplicity of antbird songs. More complex songs would be expected to require models with a higher number of states.

E. Experimental design

The number of songs recorded for each species varied from 237 for Dot-winged Antwrens to 1333 for Dusky Antbird (see Table I). The files were selected so as to contain very different amounts of background noise and highly variable song lengths, thus providing a realistic environment for testing the system. Unless otherwise noted, features were extracted using the method MFCC_0_A_D. Several HMMs design issues were investigated:

1. Number of training samples

HMMs were trained with 10, 20, 50, and 100 samples (TS) to test the relative performance of training set size. In each case, 50 recognition trials were performed. For each test there was a random repartition of the song samples into training and testing sets, with no overlap between the two.

2. Quality of training and testing samples

The effects of sample quality were investigated by using one set of 50 songs with high quality recordings. First, HMM performance was measured with these clean training samples (CTS), 10 samples for training and 40 for testing. In a second set of experiments, HMM performance was measured again, but this time using 10 and 50 CTS for training, and all the remaining A–E quality songs for testing.

3. Feature extraction methods

The performance of two different song feature extraction methods, MFCC and LPC, was investigated in order to determine which one was best for each species. Optional MFCC features included: Energy (E), Delta coefficients (D), Acceleration (A), and zeroth order (0).

4. Frequency range and number of states

The role of frequency range was addressed in a final series of experiments. In one set of experiments, features were extracted over the entire frequency range of the recording, while in another set features were extracted only in the frequency range [800–4000] Hz. The experiment was run 50 times using the method MFCC_E_D_A, with WINDOWSIZE =25 ms and TARGETRATE=15 ms, with 50 TS and the remaining samples for testing. The set of tests was repeated, this time varying the number of states in the Markov model, from 5 to 15 states.

III. RESULTS

In this section a brief summary of the obtained results is presented to illustrate several aspects of HTK for bird species recognition.

A. Number and quality of training samples

The percentage of calls correctly identified to species was almost perfect (99.5% overall recognition) when only clean recordings (grades A–C) were used for both training and testing, and only one song out of 200 test samples was misidentified even though only 10 samples per species have been used to train the models. Performance decreased slightly when the testing set included calls of all grades A–E, though recognition was still typically quite high, as seen in Fig. 7 and Table I. For both training sets, performance increased when more training samples were used, though not linearly. Looking first at the TS training sets, recognition was most successful with 100 training files (96.38% overall recognition), but was still high (93.04% overall) with just 10 training files per species. It appears that the standard deviation of performance decreased when more training sets were used, probably because more training samples captured a better approximation of the background noise. The results are substantially the same for all species, but different feature extraction methods are more appropriate for some species and less for others, suggesting that antbirds might use different features to convey information, though many more testing samples would be required to statistically test this idea.

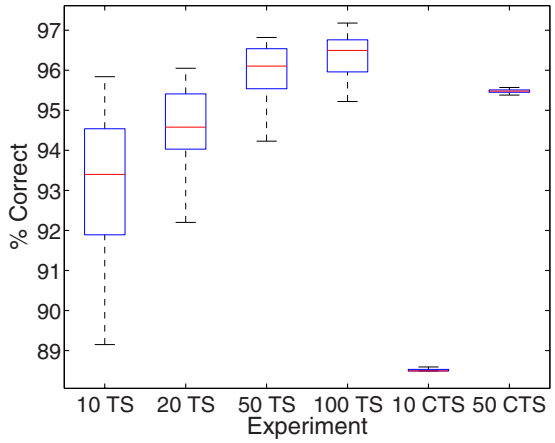

FIG. 7. (Color online) Recognition performance averaged across all species, with variable numbers of training samples. The box has lines at the lower quartile, median, and upper quartile values. The lines extending from each end of the box show the extent of the rest of the data.

Concerning the CTS training set, the overall performance was slightly lower when a large training set (50 songs per species) was used, but poorest overall (89%) with a small (10 CTS) training set. This suggests that recognition performance is highly related with the amount of background noise in the set used for training, and that data with various amounts and nature of noise during training permitted better generalization, and subsequently had better performance when tested in real environments. The standard deviation of performance in these cases is quite small because the same training data were used in every test, and a high proportion of the test songs were used in all tests, resulting in only a few independent tests.

B. Feature extraction, frequency range, and number of states

The different MFCC feature extraction methods did not differ appreciably in performance, though LPC was substantially poorer for all species (see Table II).

Given that noise affects the recognition performance, as seen in the difference between the TS and CTS trials, band-limiting is often useful to reject unwanted frequencies or avoid allocating filters to frequency regions in which there is no useful signal energy. This limitation is actually used by the MFCC method, where the frequency bands are distrib-

TABLE II. Effects of varying parameters in HTK for species recognition. The models were trained using 10 CTS. The characters after the underscore are optional features for HMM recognition: E stands for energy, D stands for delta coefficients, 0 for zeroth-order coefficients, and A stands for acceleration (the reader is invited to refer to the text for more information about the optional features).

Species	Total songs	MFCC	MFCC_E	MFCC_E_D	MFCC_0_A_D	MFCC_E_D_A	LPC
Barred Antshrike (BAS)	360	96.20	95.31	95.61	95.30	97.94	47.53
Dusky Antbird (DAB)	1333	92.76	88.12	93.68	85.84	87.98	45.28
Dot-winged Antwren (DWAW)	237	96.33	95.87	100.00	99.54	97.26	35.79
Great Antshrike (GAS)	525	80.04	74.69	82.41	81.23	77.47	28.25
Mexican Antthrush (MAT)	913	90.49	92.18	90.17	91.28	91.72	38.37
Overall performance (%)		90.78	88.42	91.60	88.50	89.03	40.37

uted only across a restricted range, instead of whole frequencies between 0 Hz up to the Nyquist frequency. As expected, the performance of the system is significantly superior when the frequency range is bounded according to the spectrum of the bird under consideration; the reason is that the noise located outside this frequency range is not taken into account by the HMMs. An average performance of 88.4% was observed when the frequency range was unrestricted, and 95.5% when the frequency was limited to the range [800–4000] Hz.

Overall performance decreased when the number of states was varied between 5 and 15 states. When 20 TS are used, performance dropped from 94.6% to 82.5%, when using six states instead of five. Adding more than 15 states degraded the performance strongly. One explanation would be that the addition of more states degrades performance because it increases the number of parameters to be estimated, thereby increasing the number of training samples that are required to obtain good generalization properties.

IV. DISCUSSION

The results presented earlier suggest that hidden Markov models are an effective method for discriminating bird species using vocalizations recorded in the field. Correct classification was superior to 95% in several experiments reported here, with 50 training samples used to discriminate between the songs of five antbird species in the tropical rainforest. When only songs with a high SNR (which is still low in comparison with laboratory recordings due to considerable ambient noise present in rainforest) were used for training and testing, only one song out of 200 was misidentified. This indicates the potential of using these methods to differentiate signals, even where noise is prevalent.

Because bird songs are modeled as a time series when using HMMs—unlike with pattern-matching methods, the length of the songs can vary without significantly affecting the recognition performance. In a Markov chain, every single observation is assumed to be statistically independent from the previous ones. Even if this is rarely the case with real signals such as bird songs or human speech, the performance was observed to be reasonably high.

Performance could be improved by developing methods to increase SNR. Recent progress in acoustic source localization using in microphone arrays allows one to implement beamforming algorithms that can enhance the quality of a signal coming from a particular direction. Approximate maximum likelihood methods that can localize antbird songs quite accurately have been developed, and experiments where beamforming is used to efficiently separate the songs of two distinct sources recorded simultaneously are presented in Chen *et al.* (2006).

HMM methods have been developed largely for understanding human speech. As a result, the most widely used software packages, such as HTK, are optimized for that application, rather than for bird songs. One example of this was observed in the experiments reported here, where LPC feature extraction performed poorly compared to MFCC. Kogan and Margoliash (1998) pointed out that LPC is appropriate to model quasilinear signals such as human voice. Such linear parametrizations do not have the ability to efficiently represent sharp transitions and other nonlinearities commonly present in bird songs. Nelson (1989) demonstrated that the dominant frequency of a call is one of the most efficient acoustic cues used by animals to convey individual information. In contrast to linear prediction coefficients, it is more likely for a bird brain to process and extract frequency information from a signal. It is possible then, that HMM models that are optimized for bird vocalizations might lead to an improved performance, for example by using specific feature extraction methods based on information gathered with data mining methods.

The methods described here were kept consistent across the five species, despite differences in signal structure. One might increase performance by introducing separate bandpass filters for every training and test signal for each species. One might also increase performance by introducing specific feature analyses that aid in species recognition from data mining methods (e.g., Vilches *et al.*, 2006). However, to run such a method in real time, it would be necessary to run the tests sequentially for one species at a time, and this for every species. In practice, the computational demands would certainly become overwhelming as the number of considered species increases.

HMMs do have certain limitations. Special care must be taken when preparing and choosing the training samples in order to obtain a high generalization ability. Performance was comparable when using 50 and 100 training samples with five antbird species, but many more samples would be required for discriminating between signals that are more similar, and particularly between songs of different individuals of the same species. Nonetheless, this method has been successfully used to identify different individuals of acorn woodpeckers.

Detecting each species occurring within a community appears to be an achievable goal. The focus in this article has been on classification. An automated species recognition system would also require event detection from streaming data with a classification system. Event detection algorithms have been developed recently using energy and entropy as the identifying criterion, e.g., (Ali *et al.*, 2007; Trifa, 2006, 2007), with varying success rates. Detected events are then sent to HTK for classification and species presence, and performance of this fully automated procedure using raw field recordings lasting several hours can be ascertained.

Combining sensor network technology with effective event detection and species identification algorithms can allow for observation of species interactions at a finer spatial and temporal scale than previously possible. Analysis of animal behavior through unattended real-time tracking of individuals and recording of large amounts of data with remote devices has become a realistic goal.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under Award No. 0410438. The authors would like to thank Edward Stabler, Yuan Yao, Yoosook Lee, Ying

Lin, and Ansgar Koene for their help and valuable comments on this paper. We also thank Martin Cody for his help with the field collections, Robert Walsh for his able assistance with the recordings, and the people of Estacion Chajul for their hospitality while we were collecting the data.

Ali, A. M., Yao, K., Collier, T. C., Taylor, C. E., Blumstein, D. T., and Girod, L. (**2007**). "An empirical study of collaborative acoustic source localization," in *Proceedings of the Sixth International Conference on Information Processing in Sensor Networks*, Cambridge, MA (ACM, New York), pp. 41–50.

Anderson, S. E., Dave, A. S., and Margoliash, D. (**1996**). "Template-based automatic recognition of birdsong syllables from continuous recordings," J. Acoust. Soc. Am. **100**, 1209–1219.

Ashiya, T., and Nakagawa, M. (**1993**). "A proposal of a recognition system for the species of birds receiving bird calls—An application of recognition systems for environmental sound," IEICE Trans. Fundamentals **E76-A**, 1858–1860.

Blumstein, D. T., and Munos, O. (**2005**). "Individual, age and sex specific information is contained in yellow-bellied marmot alarm calls," Anim. Behav. **69**, 353–361.

Catchpole, C. K., and Slater, P. J. B. (**1995**). *Bird Song: Biological Themes and Variations* (Cambridge University Press, New York).

Charif, R. A., Clark, C. W., and Fristrup, K. M. (**2006**). "Raven 1.3 user's manual," Technical Report, Cornell Laboratory of Ornithology, Ithaca, NY.

Chen, C., Ali, A., Wang, H., Asgari, S., Park, H., Hudson, R., Yao, K., and Taylor, C. E. (**2006**). "Design and testing of robust acoustic arrays for localization and beamforming," in *IEEE Proceedings of the Sixth International Conference on Information Processing in Sensor Networks*, Cambridge, MA.

Chesmore, D. (**2004**). "Automated bioacoustic identification of species," An. Acad. Bras. Cienc. **76**, 435–440.

Clark, C., Marler, P., and Beeman, K. (**1987**). "Quantitative analysis of animal vocal phonology: An application to swamp sparrow sound," Ethology **76**, 101–115.

Fagerlund, S. (**2004**). "Automatic recognition of bird species by their sounds," Master's thesis, Helsinki University of Technology, Helsinki, Finland.

Fagerlund, S., and Härmä, A. (**2004**). "Parametrization of inharmonic bird sounds for automatic recognition," in 13th European Signal Processing Conference, Antalya, Turkey, September 4–8, 2005.

Härmä, A. (**2003**). "Automatic identification of bird species based on sinusoidal modeling of syllables," in IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP 2003), April 6–10, Hong Kong, China.

Kogan, J. A., and Margoliash, D. (**1998**). "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," J. Acoust. Soc. Am. **103**, 2185–2196.

Kwan, C., Mei, G., Zhao, X., Ren, Z., Xu, R., Stanford, V., Rochet, C., Aube, J., and Ho, K. (**2004**). "Bird classification algorithms: Theory and experiemental results," in Acoustic, Speech, and Signal Processing, 2004 Proceedings, IEEE International Conference, Montreal, Canada, May 2004.

McIlraith, A. L., and Card, H. C. (**1997**). "Birdsong recognition using backpropagation and multivariate statistics," IEEE Trans. Signal Process. **45**, 2740–2748.

Nelson, D. A. (**1989**). "The importance of invariant and distinctive features in species recognition of bird song," Condor **91**, 120–130.

Rabiner, L. E. (**1989**). "A tutorial on hidden Markov models and selected applications in speech recognition," in Proc. IEEE **77**, 257-286.

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

Scheifele, P. M., Andrew, S., Cooper, R. A., Darre, M., Musiek, F. E., and Max, L. (**2005**). "Indication of a Lombard vocal response in the St. Lawrence River beluga," J. Acoust. Soc. Am. **117**, 1486–1492.

Sczewczyk, R. E., Osterweil, J., Pllastre, M., Hamilton, M., Mainwaring, A., and Estrin, D. (**2004**). "Habitat monitoring with sensor networks," Commun. ACM **47**, 34–40.

Somervuo, P. (**2000**). "Self-organizing maps for signal and symbol sequences," Ph.D. thesis, Helsinki University of Technology, Helsinki, Finland.

Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B., and Mitra, P. P. (**2000**). "A procedure for an automated measurement of song similarity," Anim. Behav. **59**, 1167–1176.

Trifa, V. (**2006**). "A framework for bird songs detection, recognition and localization using acoustic sensor networks," Master's thesis, University of California Los Angeles and École Polytechnique Fédérale de Lausanne.

Trifa, V., Girod, L., Collier, T., Blumstein, D., and Taylor, C. (**2007**). "Automated wildlife monitoring using self-configuring sensor networks deployed in natural habitats," in *Proceedings of the International Symposium on Artificial Life and Robotics (AROB 12th 2007)*, Beppu, Japan.

Vilches, E., Escobar, I. A., Vallejo, E. E., and Taylor, C. E. (**2006**). "Data mining applied to acoustic bird species recognition," in 18th International Conference on Pattern Recognition (ICPR 2006), 20–24 August, Hong Kong, China.

Wilde, M., and Menon, V. (**2003**). "Bird call recognition using hidden Markov models," Technical Report, EECS Department, Tulane University.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (**2002**). "HTK Book 3.2.1," Cambridge University Engineering Department.