Bird Species Recognition by Comparing the HMMs of the Syllables

Chih-Hsun Chou*, Chang-Hsing Lee and Hui-Wen Ni
Department of Computer Science and Information Engineering,
Chung Hua University, No.707, Sec.2, WuFu Rd.,
Hsinchu, 30067 Taiwan, R.O.C.
**chc@chu.edu.tw*

Abstract

In this study, a bird species recognition system based on their sounds is proposed. In this system, the birdsong of a bird species is segmented into many syllables, from which several primary frequency sequences can be obtained. By using the statistics of the principle frequency sequences, all the syllables are clustered with the fuzzy C-mean clustering method so that each syllable group can be modeled by a hidden Markov model (HMM) characterizing the features of the song of the bird species. Using the Viterbi algorithm, the recognition process is achieved by finding the template bird species that has the most probable HMMs matching the frequency sequences of the test birdsong. Experimental results show that the proposed system can achieve a recognition rate of over 78% for 420 kinds of bird species.

1. Introduction

There are a lot of studies of human speaker recognition, and some of them have been applied to bird species recognition. However, there is diversity in the vocalization of any particular bird species as in the case of human beings, and specific sound features are required for bird species recognition. The vocalization types of birds are birdsong and birdcall. Birdsong being complicated, varied, agreeable and pleasant to listen to is usually generated by a male bird and is used to declare his turf and attract a mate. Birdcall, on the other hand, is monotonous, brief, repeated, fixed and sexless and is used to contact or alert companions.

Birdsongs are typically divided into four hierarchical levels: note, syllable, phrase, and song [1], of which syllable plays an important role in bird species recognition. The DTW algorithm was used in a study to recognize the syllables of two bird species [2]. The authors in [3] found that many bird sounds have clear harmonic spectrum structures, and they used them to classify bird syllables into four classes. A template-based technique combining time delay neural networks was proposed to automatically recognize the syllables of 16 bird species [4]. In [5] syllables were used to deal with the overlapping problem of the sound waveforms of multiple birds, and their frequencies and amplitudes were used to form the feature vectors for recognizing 14 bird species. Instead of extracting features syllable by syllable, the histogram based on consecutive syllables was used to reveal the temporal structure of the birdsong [6]. Combination of syllables with other features can be found in [7-9].

In this study, syllables of a birdsong were extracted and used to construct the principle frequency sequence for HMM modeling in order to be able to characterize a bird species. By using the Viterbi algorithm, the recognition of a test birdsong was determined by finding the template bird species that had the biggest number of probable HMMs. After the Introduction, the developed automatic bird species recognition system is described in Section 2; Section 3 presents simulation results for examining the system performance; Section 4 gives a conclusion.

2. The proposed recognition system

The block diagram of the proposed system containing the training part and the recognition part is shown in Fig. 1. In Fig. 1, after establishing the syllable HMMs for the template bird species, the recognition process was achieved by comparing the matching degrees of the test birdsong to the HMMs of all the template birdsongs. The details are illustrated in the following.

2.1. The extraction of principle frequencies

Birdsongs are non-stationary signals requiring short-time analysis. In this study, a rectangular window with size N equaling 512 samples was applied.

0-7695-2882-1/07 $25.00 ©2007 IEEE

After segmentation, the FFT was applied to each frame with the following equation

$$X[k]=\sum_{n=0}^{N-1}x[n]\,e^{-j2\pi\frac{k}{N}n},\ k=0,1,\cdots,N-1, \quad (1)$$

which in polar coordinate form becomes

$$X[k]=\left|X[k]\right|e^{j\theta[k]}, \quad (2)$$

where k is the frequency index called frequency bin, and $\left|X[k]\right|$ and $\theta[k]$ denote the spectrum magnitude and spectrum phase of bin k.

Apply the short-time Fourier transform to each frame, which results in a magnitude spectrum and a phase spectrum. The magnitude spectrum of each frame was used to obtain a narrow time-frequency chart. Align all the time-frequency charts of a birdsong to form the time-frequency spectrogram of the birdsong signal. In this spectrogram, the grey levels reflect the strengths of the frequency components. Each trajectory represents a syllable pattern.

To extract the principle frequency of a frame, for example frame *m*, the frequency Bin in frame *m* with the greatest grey level (magnitude) was reserved and denoted by the symbol bin_m as follows:

$$bin_m=\arg\max_{0\le k\le(N-1)/2}\left[\left|X[k]\right|\right]. \quad (3)$$

This reserved frequency Bin, called the principle frequency of frame *m*, represents the spectral peak of the frame. Both the principle frequency and its corresponding magnitude form a feature vector of the frame as follows:

$$\mathbf{f}[m]=\begin{bmatrix} bin_m \\ \left|X[bin_m]\right| \end{bmatrix}. \quad (4)$$

2.2. The principle frequency sequences

Obtaining the principle frequency sequence of a syllable requires syllable segmentation from the spectrogram, as described in the following:

1. Find the feature vectors of all frames of the birdsong signal.

$$\mathbf{f}[m]=\begin{bmatrix} bin_m \\ \left|X[bin_m]\right| \end{bmatrix},\ m=1,2,\cdots,M. \quad (5)$$

2. Initialize the syllable index j, $j=1$.
3. From the feature vectors of all frames, compute the frame *t* at which the maximum magnitude occurs

$$t=\arg\max_{1\le m\le M}\left(\left|X[bin_m]\right|\right), \quad (6)$$

and set the amplitude of syllable *j* as

$$A_j=20\cdot\log_{10}\left|X[bin_t]\right|(\text{dB}). \quad (7)$$

4. Start from frame *t* and move backward and forward until frames h_j and t_j such that both $20\cdot\log_{10}\left|X[bin_{h_j}]\right|$ and $20\cdot\log_{10}\left|X[bin_{t_j}]\right|$ are smaller than $(A_j-\alpha)$. h_j and t_j are called the head frame and tail frame of syllable *j*. In this study the parameter α was set as 25 dB.
5. Record $\mathrm{B}_{s_j}=bin_{h_j}bin_{h_j+1}\ldots bin_{t_j-1}bin_{t_j}$ as the principle frequency sequence of syllable *j*.
6. Update the frame feature vectors of the birdsong by

$$\mathbf{f}[m]=\begin{bmatrix}0\\0\end{bmatrix},\quad m=h_j,h_j+1,\cdots,t_j-1,t_j. \quad (8)$$

7. Let $j=j+1$.
8. Repeat Step 3 to Step 7 until $A_j<A_1-\alpha$.

2.3. Syllable clustering

A birdsong is usually composed of many syllables but has only a few syllable patterns because of syllable repetitions. It is more practical to build an HMM for a group of similar syllables than each one of them. In this study, the fuzzy c-mean (FCM) clustering method was applied to cluster the syllables. Assume *J* syllables are obtained from the song of a bird species, denote the corresponding *J* principle frequency sequences as B_{s_j}, $j=1,2,...,J$, then the statistic vector of j^{th} syllable is computed by

$$\mathbf{v}_{s_j}=\begin{pmatrix} E[(\mathrm{B}_{s_j})^1] \\ E[(\mathrm{B}_{s_j})^2] \\ E[(\mathrm{B}_{s_j})^3] \end{pmatrix} \quad (9)$$

where $E[(\mathrm{B}_{s_j})^k]$ denotes the k^{th} moment of the principle frequencies of syllable *j*, and is calculated by

$$E[(\mathrm{B}_{s_j})^k]=\sum_{s_j=h_j}^{t_j}(bin_{s_j})^k\,\frac{1}{t_j-h_j+1}. \quad (10)$$

The FCM algorithm was applied to cluster the statistic vectors of a birdsong when the variance of the statistic vectors was greater than a predefined threshold value. In the clustering process the optimal cluster number c_{opt} was determined by using the *WB* index proposed

in [10]. The principle frequency sequences of a syllable cluster were used to establish an HMM.

2.4. Construct HMMs for the bird species

In this study, a 3-stage ergodic fully connected HMM as shown in Fig 2 was utilized. The state-transition probability matrix is denoted by $\mathbf{A}_{3\times3}$, in which a_{ij}, $i=1,2,3$, $j=1,2,3$ are nonnegative. Since the training data are the principle frequency sequences of a syllable group, the principle frequencies form the possible observations of each state in the training process. Meanwhile, the possible principle frequencies (the frequency Bins) range from 0 to 255, so that the possible observations of each state were $V=\{0,1,2,...,255\}$. So in this study, to train an HMM is to determine the initial state probabilities π_i, $i=1,2,3$, the state transition probability matrix $\mathbf{A}_{3\times3}$, and the state observation probability $B=\{b_i(v_k)\}$, $v_k=0,1,2,...,255$, $i=1,2,3$, by using the principle frequency sequences (the observation sequences) of a syllable group.

Assume there are k principle frequency sequences in a syllable group represented by $\mathbf{B}_s=\{\mathrm{B}_{s_1},\mathrm{B}_{s_2},\ldots,\mathrm{B}_{s_k}\}$, then $\mathbf{B}_s$ forms the observation sequence set for training an HMM. To train the HMM parameters by using $\mathbf{B}_s$, a well-applied expectation maximization (EM) algorithm, called a Baum-Welch algorithm, was used [11]. After the training phase, each template bird species was modeled by some HMMs. To recognize a test birdsong, after extracting its principle frequency sequences, the principle frequency sequences in the same syllable cluster were linked as a single observation sequence with which the Viterbi algorithm could be applied to find the most probable HMM λ (bird species) that generates the sequence [11]. A test birdsong signal usually contains several syllable clusters resulting in several matching degrees of the most probable HMMs. These matching degrees were used for species recognition as demonstrated in the next section.

3. Experimental results

The bird species vocalization database was obtained from a commercial CD [12], which contains both birdsongs and birdcalls of 420 bird species making it a much larger database than those used in previous studies. The sampling rate of these vocalization signals is 44.1 kHz with 16-bit resolution and a monotone type PCM format. In the experiment, the frame size was set as 512 samples with three-fourths frame overlapping. For each experiment, two-thirds of the birdsongs were randomly selected for training, and the remaining for testing. The recognition rate *RR* was defined as the number of species recognized correctly divided by the number of all species.

A. Experiments with different scales of HMMs

In the recognition phase, a test birdsong usually has several matching degrees to the most probable HMMs, requiring a rating method for recognition. Let syllable k of the test birdsong and syllable j of the template birdsong i be denoted as $test_k$ and $temp_{ij}$. The matching degree of $test_k$ with respect to the HMM is represented by $m(test_k, temp_{ij})$, $k=1,2,...,s$, where s is the syllable number of the test birdsong. In the match process, for each test syllable $test_k$, find the most likely template syllable,

$$\arg\max_{i,j} m(test_k, temp_{ij}), \tag{11}$$

then the times of i appearing in (11) for all k, $k=1,2,...,s$, denotes the number of votes for bird species i. The species with the largest number of votes is identified as the bird species of the test birdsong.

Each experiment was performed 30 times, and the performance indices included maximum recognition rate (Max), minimum recognition rate (Min), average recognition rate (Ave) and variance of recognition rate (Var). The experimental results using three scales of HMMs are given in Table 1. Table 1 shows that the system achieved average *RR*s of about 78%. On the other hand, the results show that the 3-stage HMM performed better than the other two scales, so for the following experiments, the 3-stage HMM was used.

B. Experiments with different feature dimensions

As stated in eq. (9), the first three moments of the principle frequencies in B_{s_j} were calculated to construct the three-dimensional feature vector $\mathbf{v}_{s_j}$ for syllable s_j. In this experiment, the first one, three, five and seven moments were computed to form the feature vectors with different dimensions. Each case was done 10 times for comparison as shown in Table 2. It was found that the 3-dimensional case exhibited superior results in the first three performance indices and a

comparable result for the last. This result supports the use of a 3-dimensional feature vector.

4. Conclusions

When recognizing bird species by their songs, the characteristics like width of the spectrum, energy concentration and rapid spectrum variety make the recognition process distinct from that of the human voice. In this study, the principle frequency sequences were extracted for species HMM modeling. By using the Viterbi algorithm, the recognition process could be achieved by finding the most probable HMM models for the test birdsong. An average *RR* of 78.3% was achieved by the proposed recognition system.

References

[1] C.K. Catchpole and P.J.B. Slater, *Bird song*: *biological themes and variations*, Cambridge University Press, 1995.

[2] S.E. Anderson, A.S. Dave and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *Journ. Acoust. Soc. Amer.*, vol. 100, no. 2, pp. 1209-1219, 1996.

[3] A. Härmä and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *Proc. IEEE Intern. Conf. Acoust.*, *Speech*, *Signal Proc.*, vol. 5, pp. V-701-4, 2004.

[4] S.A. Selouani, et al., "Automatic birdsong recognition based on autoregressive time-delay neural networks," in *Proc. ICSC Congr. Comput. Intellig. Methods Appli.*, pp. 1-6, 2005.

[5] A. Härmä, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Proc. IEEE Intern. Conf. Acoust., Speech, Signal Proc.*, vol. 5, pp. 545-548, 2003.

[6] P. Somervuo and A. Härmä, "Bird song recognition based on syllable pair histograms," in *Proc. IEEE Intern. Conf. Acoust., Speech, Signal Proc.*, vol. 5, pp. V-825-8, 2004.

[7] A.L. McIlraith and H.C. Card, "Bird song identification using artificial neural networks and statistical analysis," in *Proc. Canadian Conf. Electr. Comp. Engin.*, vol. 1, pp. 63-66, 1997.

[8] A.L. McIlraith and H.C. Card, "A comparison of backpropagation and statistical classifiers for bird identification," in *Proc. IEEE Intern. Conf. Neural Networks*, vol. 1, pp. 100-104, 1997.

[9] A.L. McIlraith and H.C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Trans. Signal Proc.*, vol. 45, no. 11, pp. 2740-2748, 1997.

[10] J.H. Tan, *On cluster validity for fuzzy clustering*, Master Thesis, Applied Mathem. Dep., Chung Yuan Christian University, Taiwan, R.O.C., 2000.

[11] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[12] T. Kabaya and M. Matsuda, *The Songs & Calls of 420 Birds in Japan*, SHOGAKUKAN Inc., Tokyo, 2001.

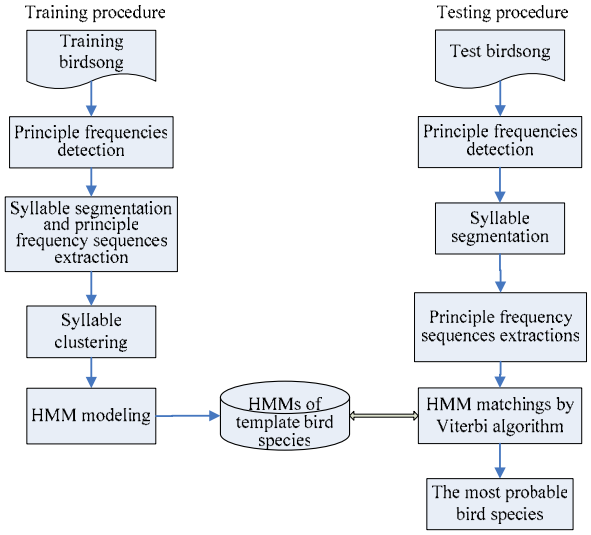

Figure 1 Block diagram of the proposed system

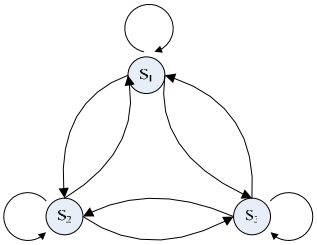

Figure 2 Applied 3-state ergodic HMM

Table 1 Experimental results with three types of HMMs.

Indices	3-state	4-state	5-state
Max	80.6	81.1	81.8
Min	74.9	74.4	72.7
Avg	78.2	78.0	77.6
Var	2.1	1.0	2.8

Table 2 Experimental results with different feature dimensions

RR(%)	1-D	3-D	5-D	7-D
Max	79.2	80.6	79.2	77.6
Min	70.9	75.6	74.9	74.7
Avg	75.6	78.3	76.8	75.9
Var	5.7	2.5	2.1	1.3