Bird Species Recognition by Wavelet Transformation of a Section of birdsong

Chih-Hsun Chou* and Pang-Hsin Liu

Department of Computer Science and Information Engineering,
Chung Hua University, No.707, Sec.2, WuFu Rd.,
Hsinchu, 30067 Taiwan, R.O.C.
*E-mail: chc@chu.edu.tw

Abstract

In this study, a method for birdsong recognition is proposed. In this method, after detecting the range of each syllable, birdsong sections containing a period of syllables were segmented. For each syllable of a birdsong section, the first five orders MFCCs were computed, and the same order MFCCs of all syllables were aligned so that wavelet transformation can be applied to compute the feature vector of the birdsong section. By using neural network as the classifier, the proposed method was applied to recognize the birdsongs of 420 bird species.

1. Introduction

The vocalization of bird species includes 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 or attract a mate. Birdcall, on the other hand, is monotonous, brief, repeated, fixed and sexless and is used to contact or alert companions. A birdcall are usually short and acoustically simple while a birdsong is longer and is composed of a succession of musical notes.

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. To recognize the syllables of two bird species, the DTW algorithm was used [2]. The idea of harmonic spectrum structures was used in [3] to classify four types of syllables. A template-based technique combining time delay neural networks (TDNNs) was proposed in [4] to automatically recognize the syllables of 16 bird species. In [5], syllables were used to solve the problem of the overlapping sound waveforms of multiple birds. A study was done [6] focusing on the histogram of consecutive syllables, called the syllable pair histogram. Similar histograms were used to construct the Gaussian prototypes of a birdsong. In [7]-[9], the number of syllables and the mean and deviation of syllable lengths were combined with other features to form the feature vectors for birdsong recognition.

Regarding to the above finds that almost all researches extracted the feature of individual syllable rather than a section of birdsong for bird species recognition. In this study, MFCC based feature extraction by using a section of birdsong containing several syllables is proposed. The proposed method not only extracted the features of individual syllables but also the variations of syllables.

The rest of this paper is divided as follows: Section 2 describes the proposed method, section 3 shows the experimental results, a conclusion is given in section 4.

2. The proposed method

The procedure of the proposed method entails preprocessing, feature extraction and recognition as described in detail in the following.

2.1 Preprocessing

The main purpose of preprocessing is to properly segment a period of syllables for feature extraction. It contains four steps: syllable endpoint detection, normalization, pre-emphasis and segmentation as described in the following.

2.1.1 Syllable endpoint detection. The detection method is described in the following.

1. Compute the short time Fourier transform of $x[n]$ with frame size $N = 512$

$$X[m,k] = \sum_{n=0}^{N-1} x[n] w_m[n] e^{-j2\pi mk/N}, 0 \le k < N,$$

where m is the frame number. The Hamming window for short time analysis has the form of

978-0-7695-3737-5/09 $25.00 © 2009 IEEE
DOI 10.1109/UIC-ATC.2009.85

$$w_m[n] = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right), (m-1)N \le n < mN \\ 0 \qquad\qquad\qquad\qquad\quad , \text{otherwise} \end{cases}$$

2. Form the spectrogram of the signal by aligning the spectrum of all frames, $X[m,k]$, m = 1, 2, …, M, where M is the number of frames of $x[n]$.
3. For each frame m, find the frequency Bin bin_m with the greatest magnitude.

$$bin_m = \arg \max_{0 \le k \le N-1} (|X[m,k]|), m = 1, 2, \ldots, M.$$

4. Initialize the syllable index j, $j = 1$.
5. Compute the frame t at which the maximum magnitude occurs

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

and set the amplitude of syllable j as

$$A_j = 20 \cdot \log_{10} |X[m, bin_t]| (\text{dB}).$$

6. Start from frame t and move backward and forward up to frames h_j and t_j such that both $20 \cdot \log_{10} |X[h_j, bin_{h_j}]|$ and $20 \cdot \log_{10} |X[t_j, bin_{t_j}]|$ are smaller than $(A_j - 20)$ (dB). Then h_j and t_j are called the head frame and tail frame of syllable j.
7. Set

$$|X[m, bin_m]| = 0, m = h_j, h_j + 1, \cdots, t_j - 1, t_j$$

8. Let $j = j + 1$.
9. Repeat Step 6 to Step 9 until $A_j < A_1 - 20$.

By using the above procedure, the boundaries of each syllable can be obtained. In Section 4, a comparison of birdsong recognition efficiency for both methods is given.

2.1.2 Normalization and pre-emphasis. The normalization process regulates the discrepancy of the voice amplitudes caused by the diversity of recording environments. In this study, the amplitudes were linearly normalized to the region of [-1, 1]. Besides, since the amplitude of the high frequency is usually much smaller than that of the low frequency, pre-emphasis was applied to intensify the signal of the high frequencies. The intensification was accomplished by a finite impulse response (FIR) filter with the following form

$$H(z) = 1 - a \cdot z^{-1},$$

so that a signal $x[n]$ after the filtering process has the following property

$$\bar{x}[n] = x[n] - ax[n-1],$$

in which a is a scalar usually between 0.9 and 1, which was set as 0.95 in this study.

2.1.3 Segmentation. Focusing on the recognition of a section of birdsong, the segmentation process in this system aimed at segmenting a period of syllables rather than an individual syllable similar to many other researches [2]-[4], [7]-[9]. Extracting the feature vector of a period of syllables is practical for birdsong recognition because the syllables of a birdsong are usually repetitive. After endpoint detection, normalization and pre-emphasis, the segmentation process was done by detecting the repetition of syllables as described in the following

1. Set $i = 1$ the index of the first syllable of the segmentation.
2. Find the nearest syllable j such that the similarity between syllables i and j, sim_{ij}, is smaller than α, j is the last syllable of the segmentation.
3. Set the segmentation length $l = j$.
4. Set $k = j + 1$, the checking index.
5. Set $i = 1$, $l = j$.
6. Compute the similarity, sim_{ki}, between syllable k and syllable i.
7. If $sim_{ki} > \alpha$ (the same type), then
   If $l = k - j$ then Stop. The segmentation is from syllable 1 to syllable l.
   If $i = j$, then $j = j + 1$ go to Step 5.
   set $i = i + 1$ and $k = k + 1$, and go to Step 6.
8. If $i = j$, then $j = j + 1$ go to Step 5.
9. Set $k = k + 1$, $j = j + 1$, $l = l + 1$ and go to Step 6.

The similarity between two syllables was determined by computing the difference between the amplitudes of corresponding frequency Bins. Since the syllable types of a birdsong are usually within 6, in our experiment, α was set as a value such that $2 \le l \le 8$. In fact, syllables of the same type in a section of a birdsong usually have small variances in amplitude and length. After segmentation, the segmented syllables were aligned for feature extraction.

2.2 Feature extraction – the WMFCCs

After segmenting a period of syllables, the aligned

syllables were used to compute the feature vector of the birdsong. The process for obtaining the feature vector WMFCCs is shown in Fig. 2.1 and is described below.

2.2.1 Computing the MFCCs of each frame. The steps for computing the MFCCs of each frame are as follows:

1. Compute the fast Fourier transform (FFT) of each framed signal.

$$\tilde{x}[k]=\sum_{n=0}^{N-1}x[n]w[n]e^{-j2\pi nk/N},\ 0\le k<N.$$

2. Compute the energy of each triangular filter band

$$E_j=\sum_{k=0}^{N/2-1}\phi_j[k]|\tilde{x}[k]|^2,\ 0\le j<J,$$

where $\phi_j[k]$ denotes the amplitude(weight) of the j^{th} triangular filter at frequency bin k as shown in Fig. 2.2, E_j denotes the energy of j^{th} filter band, and J is the number of triangular filters.

3. Compute the MFCCs by Cosine transformation

$$c_i(m)=\sum_{j=0}^{J-1}\cos\left(m\frac{\pi}{J}(j+0.5)\right)\log_{10}(E_j),$$

where $c_i(m)$ denotes the m^{th} order MFCC of the i^{th} frame.

Although MFCCs have been well-applied in human voice recognition, the process for birdsong recognition needs to be improved. A well-known technique called wavelet transform is adequate to tackle the resolution problem. In the following, an improvement of the MFCCs, called the WMFCCs, is obtained to form the feature vector.

2.2.2 Feature vector formed by using the WMFCCs. After obtaining the MFCCs of each frame of the aligned birdsong signal, the feature vector of the birdsong was obtained by computing the WMFCCs as described in the following.

1. Collect the MFCCs of all frames of the aligned signal.

$$\{c_1(0),c_1(1),...,c_1(L-1),...,c_i(0),...,c_i(L-1),...\},$$

where L is the total order of the MFCCs.

2. Align the MFCCs of the same order.

$$\mathbf{s_m}[n]=[c_1(m),c_2(m),...,c_i(m),...],\ \ m=0,...,L-1.$$

3. Compute 3-level wavelet transformations of $\mathbf{s_m}[n]$ as shown in Fig. 2.3. The transformation equations are

$$\mathbf{a}[n]=\sum_{k=-\infty}^{\infty}\mathbf{h_0}[k]\mathbf{s_m}[2n-k],$$

$$\mathbf{d}[n]=\sum_{k=-\infty}^{\infty}\mathbf{h_1}[k]\mathbf{s_m}[2n-k],$$

where $\mathbf{a}[n]$ and $\mathbf{d}[n]$ denote the low frequency and high frequency components of $\mathbf{s_m}[n]$, and $\mathbf{h_0}[n]$ and $\mathbf{h_1}[n]$ are the applied low pass and high pass filters in the transformation. The coefficients of these two filters are [10], [11]

$$\mathbf{h_0}[n]=[0.3327,0.8069,0.4599,-0.1350,-0.0854,0.0352],$$
$$\mathbf{h_1}[n]=[0.0352,0.0854,-0.1350,-0.4599,0.8069,-0.3327]$$

The resulted six sequences of the transformation, called the WMFCCs, are denoted as s_m^{LLL}, s_m^{LLH}, s_m^{LH}, s_m^{HL}, s_m^{HHL} and s_m^{HHH}.

4. Compute the means of each of the six sequences and denote them as ms_m^{LLL}, ms_m^{LLH}, ms_m^{LH}, ms_m^{HL}, ms_m^{HHL} and ms_m^{HHH}. In audio signal classification, a widely used feature is the means of each order of MFCCs. In this study, the means of the six sequences were computed.

5. Form the feature vector by using the six mean values of all the first five order MFCC sequences

$$[ms_0^{LLL}\ \ldots\ ms_0^{HHH}\ ms_1^{LLL}\ldots ms_1^{HHH}\ \ldots ms_4^{LLL}\ldots ms_4^{HHH}]^T.$$

Such a feature vector has a fixed length no matter the length of the birdsong section.

By using the feature vectors obtained in Step 5, the recognition process was achieved by applying the back-propagation neural network (BPNN) classifier as described in the next section.

2.3 Recognition by using the BPNN

In the BPNN training process, the feature vector of the training syllable was used as the input and the corresponding bird species as the desired output. The number of nodes in the input layer is equal to the dimension of the training vector, while the number of nodes in the output layer is equal to the number of bird species. The number of nodes in the hidden layer was set as the average of the other two layers, which is frequently adopted. A sigmoid type activation function was used for both hidden and output nodes.

In the recognition process, the feature vector of a

test birdsong was obtained by the same process as the training part. After inputting the feature vector to the BPNN, the output of the network indicated the species class the test birdsong belonged to.

3. Experimental results

The bird species vocalization database used in this study was obtained from a commercial CD [12] containing both birdcall and birdsong files of 420 bird species recorded in the field in Japan. Each file contained vocalizations of the same bird species. The database of 420 bird species made it much larger than any other used in previous studies. Meanwhile, recordings in the field were usually in a noisy environment, incomplete and interrupted. The sampling rate of these vocalization signals was 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. In the experiment, half of each birdsong file was used for training and the remaining for testing.

3.1 The MFCC-based feature vector for comparison

For the purpose of checking the efficiency of the proposed method, a MFCC-based feature vector was constructed to compare the recognition rate. To construct such a feature vector, after syllable segmentation, the MFCCs of each frame of a training syllable were obtained by using the steps in sec. 2.2.1. After computing the first 15 order MFCCs of each frame, the coefficients of the same order of all frames were averaged and then used to form a 15-dimensional syllable feature vector. Such a feature vector had been successfully used in many studies, so it was used for checking.

3.2 Experimental results

In many studies, the first fifteen order MFCCs have been used to form the feature vector. In the proposed WMFCC method, for reducing the computation complexity, only the first five orders were used. In the experiment, half of each birdsong file was used for training and the remaining for testing. By using the segmentation process, both training and testing data sets contained several birdsong sections of each bird species. When MFCC-based method was applied, all the syllables in a birdsong section (training or testing) were treated individually. In the recognition of a birdsong section, the species was determined by finding the species class with the largest number of NN output. For WMFCC approach, the feature vector of a period of syllables in the birdsong section (training or testing) was used as the input of NN. In the recognition process, the NN output indicated the species class of the test birdsong section. In this experiment, the recognition rate *RR* was defined as

$$RR(\%) = \frac{\text{number of test birdsong sections recognized correctly}}{\text{number of test birdsong sections}} \cdot 100\%$$

The *RRs* of both methods are shown in Table 3.1. It shows that the proposed WMFCC feature vector improved the *RRs* of 18.72% on the MFCCs based. The above comparisons exhibited the efficiency of the proposed feature extraction method.

4. Conclusion

In the study of birdsong recognition, Most of the presented studies used the features of individual syllables to form the feature vector of each bird species. In this study, rather than using individual syllables, a section of birdsong containing a period of syllables was used to extracting the feature vector. Experimental results showed that the proposed methods evidently improved the *RRs* over the traditional MFCC based approach.

5. 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," *Journal of the Acoustical Society of America*, vol. 100, no. 2, pp. 1209-1219, 1996.

[3] A. Härmä and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *Proceedings of the IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, vol. 5, pp. V-701-4, 2004.

[4] S.A. Selouani, M. Kardouchi, E. Hervet and D. Roy, "Automatic birdsong recognition based on autoregressive time-delay neural networks," in *Proceedings of the ICSC Congress on Computational Intelligence Methods and Applications*, pp. 101-106, 2005.

[5] A. Härmä, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing*, vol. 5, pp. 545-548, 2003.

[6] P. Somervuo, "A. Härmä, Bird song recognition based on syllable pair histograms," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal*

Processing, 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 *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, 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 *Proceedings of IEEE International Conference on Neural Networks*, vol. 1, pp. 100-104, 1997.

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

[10] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, pp. 909-996, Oct. 1988.

[11] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, 1992.

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

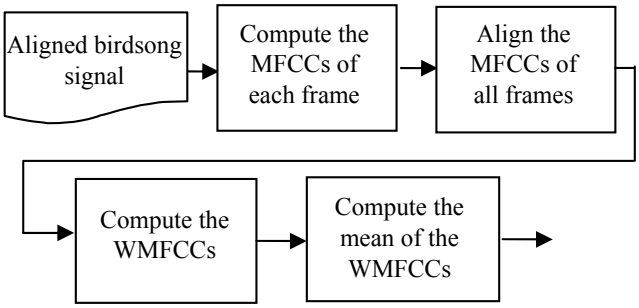

Figure 2.1 Flow chart of computing the feature vector WMFCCs

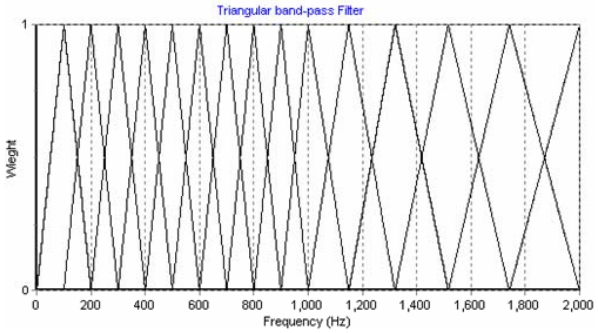

Figure 2.2 Applied triangular filters for computing the MFCCs

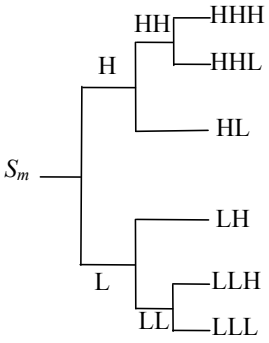

Figure 2.3 Applied 3-level wavelet transformation.

Table 3.1 *RR*s of using various structures

Structure	***RR***
MFCC+BPNN	54.69%
WMFCC+BPNN	73.41%