Automatic Recognition of Bird Songs Using Cepstral Coefficients

Article · January 2006

3 authors, including:

Chang-Hsing Lee
Chung Hua University
54 PUBLICATIONS **1,052** CITATIONS

Journal of Information Technology and Applications
Vol. 1 No. 1, May, 2006, pp.17-23

Automatic Recognition of Bird Songs Using Cepstral Coefficients

Chang-Hsing Lee
Deptartment of Computer Science and Information Engineering
Chung Hua University, Hsinchu 300, Taiwan

Yeuan-Kuen Lee
Deptartment of Computer Science and Information Engineering
Ming Chuan University, Taoyuan 333, Taiwan

Ren-Zhuang Huang
Deptartment of Computer Science and Information Engineering
Chung Hua University, Hsinchu 300, Taiwan

Abstract

In this paper we propose a method to automatically identify birds from the sounds they generate. First, each syllable corresponding to a piece of vocalization is segmented. For each syllable, the averaged LPCCs (ALPCC) and averaged MFCCs (AMFCC) over all frames in a syllable are calculated as the vocalization features. Linear discriminant analysis (LDA) is exploited to increase the classification accuracy at a lower dimensional feature vector space. In our experiments, AMFCC usually outperforms ALPCC. If a codebook consisting of several representative feature vectors is used to model the syllables of the same bird species, the average classification accuracy is 87% for the recognition of 420 bird species.

Keywords: birdsong recognition, linear discriminant analysis, LPCC, MFCC

1. Introduction

In the daily life, we can hear a variety of creature's sounds, including human speech, dog barks, bird songs, cicada sounds, frog calls, and cricket calls, etc. Many animals generate sounds either for communication or as a by-product of their living activities such as eating, moving, or flying, etc. Identification of animals by their sounds is valuable for biological research and environmental monitoring applications, especially in detecting and locating animals. In general, people often hear the sounds generated by animals rather than see the animals. Furthermore, most of the animal vocalizations have evolved to be species-specific. Therefore, the utilization of animal vocalizations to automatically identify animal species is a natural and adequate way to ecological censusing, environment monitoring, biodiversity assessment, etc.

Bird songs are usually more variable and can be regarded as "polytonic". That is, the types of sounds that birds generate and the syntactical arrangements of those sounds change significantly. Thus, bird songs are typically divided into a set of hierarchical acoustic structures [1]. The different structural components of bird songs can be described in order of increasing complexity. The simplest individual sounds that birds produce are referred to as song ''elements'' or ''notes''. A set of one or more elements that occur successively in a regular pattern is referred to as a song ''syllable''. A sequence of one or more syllables that occurs repeatedly is regarded as a song ''motif'' or ''phrase''. A particular combination of motifs that occur repeatedly constitutes a song ''type''. Finally, a sequence of one or more motifs separated from other motif sequences by silent intervals of different duration is a song ''bout''.

Anderson et al. [2] used DTW for automated analysis of continuous recordings of animal vocalizations. They directly compared signal spectrograms, and identify constituents and constituent boundaries. The feature vectors are derived from the log magnitudes of FFT bins from 0.5 to 10 KHz. They evaluated the performance on the vocalizations of an indigo bunting (*Passerina cyanea*) and a zebra finch (*Taeniopygia guttata*). The test data is collected from a low-cutter, low-noise environment. The representative templates (syllables) are segmented manually by the investigator. They identify syllables in stereotyped songs and calls with greater than 97% accuracy. Syllables in the more variable and lower amplitude plastic songs are identified with approximate 84% accuracy.

Kogan and Margoliash [3] compared two techniques, DTW and HMM, for automatic recognition of bird song elements from continuous recordings. The feature vectors used for DTW classification are the log magnitudes of FFT bins from 0.5 to 10 KHz. Six types of parameters are compared on HMM performance, including LPC, LPCCs, LPC reflection, MFCCs, log mel-filter bank channel, and linear mel-filter bank channel. Experiments show that DTW-based technique gives excellent to satisfactory performance. However, DTW requires careful selection of templates that may

need more expert knowledge for noisy recordings or presence of confusing short-duration calls. Better performance of HMM can be achieved based on segmentation and labeling of constituent vocalizations. One disadvantage of HMM is the misclassification of short-duration vocalizations or song units with more variable structure.

McIlraith and Card [4-7] had proposed several methods for birdsong recognition. Neural networks and statistical methods were used to recognize six bird songs (song sparrow, fox sparrow, marsh wren, sedge wren, yellow warbler, and red-winged blcakbird). Temporal measurements as well as spectral information are used as features in their study. The temporal measurements include the number of elements, the mean and variance of element lengths, and the mean and variance of silence lengths within each song. The spectral information includes LPC cepstral coefficients [6] or means and standard deviations of the power spectral density for nine spectral bands [7-9]. Quadratic discriminant analysis is exploited to boost the classification accuracy. The classification accuracy is 82% using backpropagation neural network for identification and 93% using quadratic discriminant analysis.

A. Harma [8] proposed a method for automatic identification of bird species based on sinusoidal modeling of syllables. For many songbirds, a large class of syllables can be approximated as amplitude- and frequency-varying brief sinusoidal pulses. A segmentation algorithm is proposed to divide a song into a set of syllables. A weighted sum of the mean differences between frequency and amplitude trajectories is evaluated for recognition purpose. Experimental results show that with limited sets of bird species a recognizer based on this signal model may be sufficient.

In this paper, the cepstral coefficients were calculated for the recognition of bird songs. For each syllable segmented from the bird songs, the averaged LPCCs (ALPCC) and averaged MFCCs (AMFCC) over all frames in the syllable were calculated as the vocalization feature. A codebook consisting of several template features is exploited to model the variant characteristics of different syllables segmented from the same bird songs. In the following section, we will describe the proposed bird song recognition method. Experimental results will be presented in Section 3 to show the effectiveness of the proposed method. Finally, a conclusion is given in Section 4.

2. The Proposed Bird Song Recognition Method

The proposed bird song recognition system consists of two phases: the training phase and the recognition phase. The training phase is composed of four main modules: syllable segmentation, feature extraction, codebook generation, and linear discriminant analysis (LDA). The recognition phase consists of four modules: syllable segmentation, feature extraction, LDA transformation, and classification. Fig. 1 shows the block diagram of the proposed bird song recognition system. A detailed description of each module will be described below.

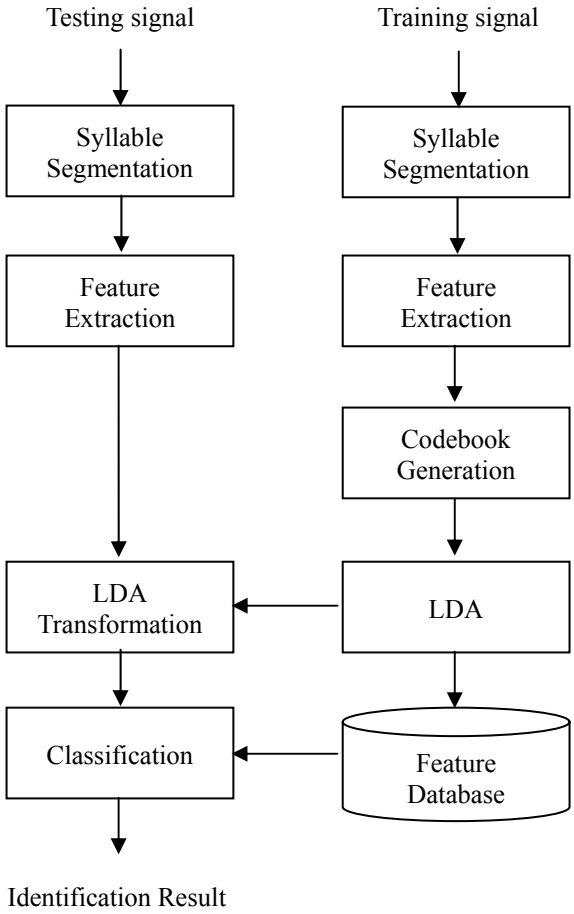

Fig. 1 The block diagram of the proposed bird song recognition system.

2.1 Syllable Segmentation

Each input bioacoustic signal is first segmented into a set of syllables using the method proposed by Harma [8]. Each syllable is regarded as the elementary acoustic recognition unit since it is relatively easier to extract a number of syllables from a recording with many animals singing simultaneously. In addition, the features extracted from the syllable are more invariant to regional variations in animal vocalizations. The detail of the syllable segmentation method is described as follows:

Step 1. Compute the spectrogram of the input bioacoustic signal using short-time Fourier transform (STFT). We denote the spectrogram a matrix $M(f, t)$, where f and t are the frequency and frame indexes, respectively.

Step 2. Set $n = 0$.

Step 3. Find f_n and t_n, such that $|M(f_n, t_n)| \geq |M(f, t)|$ for every pair of (f, t) and set the position of the n-th syllable to be t_n.

Step 4. Compute the amplitude $A_n(0) = 20\log_{10}|M(f_n, t_n)|$ dB. If $A_n(0) < A_0(0) - ß$ dB, stop the segmentation process, where $ß$ is the stopping criteria and its default value is 20. This means that the amplitude of the n-th syllable is too small and hence no more syllables need to be extracted.

Step 5. Starting from t_n, trace the maximal peak of $|M(f, t)|$ for $t < t_n$ until $A_n(t - t_n) < A_n(0) - ß$ dB, where $A_n(t - t_n) = 20\log_{10}|M(f, t)|$ dB. Next, trace the maximal peak of $|M(f, t)|$ for $t > t_n$ until $A_n(t - t_n) < A_n(0) - ß$ dB. This step is to determine the starting time $(t_n - t_s)$ and the ending time $(t_n + t_e)$ of the n-th syllable around t_n.

Step 6. Set $M(f, t) = 0$ for $t \in \{t_n - t_s, \ldots, t_n + t_e\}$ to delete the area of the n-th syllable. Set $n = n+1$ and go to **Step 3** to find the next syllable.

2.2. Feature Extraction

After each syllable is segmented from the input bird song, each syllable is then divided into a set of overlapped frames. Since birdsong and human speech share some common themes and mechanisms [14], it is postulated that the features widely used for speech recognition, such as LPCCs and MFCCs, will yield satisfactory outcome. Thus, the LPCCs/MFCCs are computed as the feature vector of each frame. The averaged LPCCs and averaged MFCCs over all frames within a syllable are calculated as the representative feature vector of the syllable. The use of the cepstral coefficients allows for the similarity between two cepstral feature vectors to be computed as a simple Euclidean distance.

2.2.1 LPCCs

Linear predictive coding (LPC) based representation of speech signals, especially LPC derived cepstrum coefficients (LPCCs), is a very effective representation for speech coding, analysis, synthesis, and recognition [11]. LPC analysis tries to find the filter characteristics that model the human vocal tract. The basic idea behind the LPC model is that a speech sample can be approximated with a linear combination of previous p speech samples:

$$\hat{s}[n] = \sum_{k=1}^{p} a_k s[n-k], \tag{1}$$

where p is called the order of the model. The filter characteristics of the human vocal tract are determined from the set of linear prediction coefficients (LPCs). The set of LPCs is determined by finding the set of parameters $\{a_k\}$ that minimizes the difference between the actual value and the predicted value over the whole set of speech samples. The spectrum based on the LPC model is defined as

$$S(e^{j\omega}) = \frac{\sigma^2}{1 - \sum_{k=1}^{p} a_k e^{-j\omega k}}, \tag{2}$$

where σ^2 is the gain term in the LPC model. A significant property of the LPC spectral modeling is that the LPC spectrum matches the signal spectrum closely near the spectral peaks. In fact, the linear predication cepstral coefficients (LPCCs) are more robust and reliable features for speech recognition and have been proven to be more relevant than LPCs [9]. The set of LPCCs $\{c_k\}$ are the Fourier representation of the log-magnitude spectrum:

$$\log_{10}\left|S(e^{j\omega})\right| = \sum_{k=-\infty}^{\infty} c_k e^{-j\omega k}. \tag{3}$$

The coefficients $\{c_k\}$ can be obtained directly from $\{a_k\}$ through the following equations:

$$c_0 = \ln \sigma^2, \tag{4}$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \leq m \leq p, \tag{5}$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad m > p. \tag{6}$$

2.2.2 MFCCs

The *mel scale* is a means of mapping the physical frequency to the perceptual representation. The mapping between the physical frequency scale (Hz) and perceptual frequency scale (*mel*) is approximately linear below 1000 Hz and logarithmic at higher frequencies. The relation between the physical frequency scale and the *mel* frequency scale can be described as:

$$mel = 2595\log_{10}(1 + \frac{f}{700}). \tag{7}$$

Mel-frequency cepstral coefficients (MFCCs) exploit the model of the human auditory system as well as the decorrelating property of the cepstrum [9-11]. MFCCs have been the most widely used features for speech recognition due to their ability to represent the speech spectrum in a compact form. In fact, MFCCs have been proven to be very effective in automatic speech recognition and in modeling the subjective frequency contents of audio signals. In the process of MFCCs extraction, the Fourier spectrum is filtered by a set of *mel*-scale filters. The MFCCs are computed by performing DCT on the logarithmic energy output of every bandpass filter:

$$c_m = \sum_{k=0}^{K-1} \log_{10}(E_k) \cos(m\frac{\pi}{K}(k+0.5)), \quad 0 \le m \le L\text{-}1, \tag{8}$$

where K is the number of bandpass filters, L is the desired length of MFCCs, and E_k is the energy of the output of the k-th bandpass filter.

2.2.3 Averaged LPCCs and Averaged MFCCs

The LPCCs/MFCCs of each frame are computed and regarded as the feature vector of this frame. Since the number of frames varies for syllables with different lengths. To deal with this problem, the averaged LPCCs/MFCCs over all frames in a syllable are used as features. Therefore, the number of feature values is fixed regardless of the length of the acoustic syllable. The averaged feature value of a syllable is computed by

$$C_m = \frac{1}{M}\sum_{i=1}^{M} c_m^i \ , \ 0 \le m \le L\text{-}1, \tag{9}$$

where M is the number of frames in the syllable, L is the dimension of the feature vector, C_m is the m-th feature value, and c_m^i denotes the m-th cepstral value of the i-th frame.

In the training phase, the average of the feature values of all the training syllables segmented from the same species is regarded as the feature value of this species:

$$F_m = \boldsymbol{E}(C_m), \ 0 \le m \le L\text{-}1, \tag{10}$$

where $E(\cdot)$ denotes the expected value (statistically average). Since the dynamic range differs for each feature value, a linear normalization is applied to obtain the normalized feature value, f_m:

$$f_m = \frac{F_m - F_m^{\min}}{F_m^{\max} - F_m^{\min}}, \ 0 \le m \le L\text{-}1, \tag{11}$$

where $F_m^{\max}$ and $F_m^{\min}$ denote respectively the maximum and minimum of the m-th feature value for all training syllables.

2.3 Codebook Generation

As mentioned previously, the bird songs are "polytonic". That is, each bird song consists of a set of syllables with variant characteristics. In fact, the feature vectors of two syllables segmented from the same song may have significant difference. Therefore, it is better to model the same bird song with a codebook consisting of several feature vectors. That is, a set of feature vectors is used to model the same bird song. The set of representative feature vectors is automatically calculated from the set of feature vectors extracted from the training syllables. A clustering algorithm, called progressive constructive clustering (PCC) [12], is used to automatically divide the set of training feature vectors belonging to the same bird species into several subcategories. Each subcategory consists of a set of similar feature vectors and thus will be represented by the mean feature vector of the same subcategory. Let $S^j = \{\mathbf{x}_1^j, \mathbf{x}_2^j, \cdots, \mathbf{x}_{N_j}^j\}$ be the set of training feature vectors for the j-th bird species, where N_j is the cardinality of S^j. A detailed description of the PCC algorithm is described as follows:

Step 1. Take $\mathbf{x}_1^j$ as the feature vector of subcategory 1, $\mathbf{f}_1$. Set $i = 2$ and $nc = 1$.

Step 2. Take $\mathbf{x}_i^j$ and find its nearest feature vector $\mathbf{f}_k$ from the set $\{\mathbf{f}_1, \ldots, \mathbf{f}_{nc}\}$. Calculate the distance between $\mathbf{x}_i^j$ and $\mathbf{f}_k$, $d(\mathbf{x}_i^j, \mathbf{f}_k)$. If $d(\mathbf{x}_i^j, \mathbf{f}_k) \le T_d$, $\mathbf{x}_i^j$ will be assigned to subcategory k; otherwise, go to **Step 4**.

Step 3. Recalculate the feature vector of subcategory k where $\mathbf{x}_i^j$ is taken into consideration. Go to **Step 5**.

Step 4. Set $nc = nc + 1$ and the new class $\mathbf{f}_{nc} = \mathbf{x}_i^j$.

Step 5. If $i = N_j$, exit; otherwise, set $i = i + 1$ and go to **Step 2**.

2.4 Linear Discriminant Analysis (LDA)

LDA [11] aims at improving the classification accuracy at a lower dimensional feature space. LDA deals with discrimination between classes rather than representations of various classes. The goal of LDA is to minimize the within-class distance while maximizing the between-class distance. In LDA, an optimal transformation matrix from an n-dimensional feature space to d-dimensional space is determined, where $d \le n$. The most widely used transformation matrix is a linear mapping that maximizes the so-called Fisher criterion J_F:

$$J_F(A) = tr((A^T S_W A)^{-1}(A^T S_B A)), \tag{12}$$

where S_W and S_B are called the within-class scatter matrix and between-class scatter matrix, respectively. The within-class scatter matrix is defined as:

$$S_W = \sum_{j=1}^{NC}\sum_{i=1}^{N_j}(\mathbf{x}_i^j - \boldsymbol{\mu}_j)(\mathbf{x}_i^j - \boldsymbol{\mu}_j)^T, \tag{13}$$

where $\mathbf{x}_i^j$ is the i-th feature vector in class j, $\boldsymbol{\mu}_j$ is the mean vector of class j, NC is the number of classes, and N_j is the number of feature vectors in class j. The between-class scatter matrix is given by:

$$S_B = \sum_{j=1}^{NC}(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T, \tag{14}$$

where $\boldsymbol{\mu}$ is the mean vector of all classes. From Eq.

(12), we can see that LDA tries to find a transformation matrix that maximizes the ratio of between-class scatter to within-class scatter in a lower-dimensional space. The optimal solution is the $n \times d$ transformation matrix, A_{opt}, given by

$$A_{\text{opt}} = \arg\max_{A} \frac{tr(A^T S_B A)}{tr(A^T S_W A)}. \quad (15)$$

The optimal transformation matrix A_{opt} can be determined by finding the eigenvectors of $S_W^{-1} S_B$. The columns of A_{opt} are the d eigenvectors corresponding to the d largest eigenvalues. Assuming that the eigenvalues are sorted in a non-increasing order, the number of eigenvectors retained is determined by finding the minimum d such that

$$\sum_{i=1}^{d} \lambda_i \geq 0.95 \sum_{i=1}^{n} \lambda_i , \quad (16)$$

where λ_i is the i-th eigenvalue.

After deriving the optimal transformation matrix A_{opt}, we use A_{opt} to transform each normalized n-dimensional feature vector to a d-dimensional vector. Let $\mathbf{x}_j$ denote the n-dimensional feature vector of the j-th class, the reduced d-dimensional feature vector can be computed by:

$$\mathbf{f}_j = A_{opt}^T \mathbf{x}_j \quad (17)$$

2.5. Recognition Phase

At the recognition phase, each input signal is first segmented into a set of syllables. The averaged LPCCs/MFCCs for each syllable are calculated. The same linear normalization process is applied to each set of LPCCs/MFCCs. The normalized LPCCs/MFCCs are transformed to be a lower-dimensional feature vector by using the transformation matrix A_{opt}. The distance between the test feature vector and every representative feature vector is calculated by using the Euclidean distance. Let $\mathbf{x}^r$ denote the representative feature vector that has minimum Euclidean distance to the input feature vector $\mathbf{x}$:

$$d(\mathbf{x}, \mathbf{x}^r) \leq d(\mathbf{x}, \mathbf{x}^k), \ 1 \leq k \leq N, k \neq r, \quad (18)$$

where N is the number of representative feature vectors in the database. The subject code s that represents the identified species is determined by the set to which the representative feature vector $\mathbf{x}^r$ belongs:

$$s = i \ \text{ if } \ x^r \in G^i, \ 1 \leq i \leq N_s, \quad (19)$$

where G^i denote the set consisting of representative feature vectors for the i-th species and N_s is the number of species class in the database.

3. Experimental Results

The test audio signals are derived from commercially available compact disk and the sampling frequency is 44100 Hz with each sample digitized in 16 bits. In the experiments, two bird-song databases are tested. The first database consists of 420 bird songs with each song corresponding to each bird species. Bird songs for the same bird species may be further divided into two or more songs depending on their functions or characteristics. In total, the second database consists of 561 different types of bird songs. Each audio signal is first segmented into a set of syllables, half of which is used for training and the other half for testing. The classification accuracy (CA) is defined as:

$$CA = \frac{N_{CA}}{N_S} \times 100\%, \quad (20)$$

where N_{CA} is the number of syllables which were recognized correctly and N_S is the total number of test syllables.

Two experiments are conducted in which only the averaged feature vector or several representative feature vectors were employed to identify each species. In the experiments, two different features, LPCCs and MFCCs, are conducted to compare their performance. In addition, LDA is exploited to reduce the feature dimension and improve the classification accuracy. The number of the cepstral coefficients is 15. That is, the length of LPCCs/MFCCs is 15.

3.1 Experiment 1

Table 1 shows the classification accuracy of HMM, ALPCC and AMFCC with/without LDA, where B420 and B561 denote the two databases consisting of 420 and 561 bird songs, respectively. From this table, we can see that AMFCC greatly outperforms HMM and ALPCC. HMM, which exploits the temporal information among the frames within a syllable, does not provide better performance than ALPCC or AMFCC. Since variations between consecutive frames within a syllable are not regular and thus temporal information extracted for identification purpose is not very essential. In addition, most of the sounds are recorded in the field with additional sounds/noise in the background. Thus, the feature vector extracted from each syllable is not stable enough. On the other hand, the averaged LPCC/MFCC can attenuate the effect of background noise by averaging the feature vectors of all frames within the syllable. Therefore, the proposed AMFCC is adequate for the identification of birdsongs. In addition, the performance can also be improved with LDA. Table 2 shows the reduced feature dimension after applying LDA. The classification accuracy can be improved if the bird songs are first manually

classified into different types according to their functions.

Table 1
The classification accuracy of ALPCC and AMFCC with/without LDA.
(T_N is the number of testing syllables)

	HMM	ALPCC	ALPCC +LDA	AMFCC	AMFCC +LDA	T_N
B420	26%	26%	39%	50%	57%	14733
B561	28%	36%	47%	58%	68%	14621

Table 2
The reduced feature dimension after applying LDA.

	ALPCC	ALPCC +LDA	AMFCC	AMFCC +LDA
B420	15	11	15	11
B561	15	11	15	11

3.2 Experiment 2

As mentioned previously, the characteristics of syllables taken from the same bird songs may vary significantly. Therefore, we use a codebook consisting of several representative feature vectors to model the syllables segmented from the same bird songs. Table 3 shows the classification accuracy of codebook-based LPCCs (CLPCC) and codebook-based MFCCs (CMFCC) with/without LDA. Table 4 shows the reduced feature dimension after applying LDA. Comparing Tables 1 and 3, we can see that the classification accuracy is boosted if each bird song is modeled by a codebook consisting of several representative feature vectors instead of only the mean feature vector. However, the computation time increases as well since the number of representative feature vectors increases.

Table 3
The classification accuracy of CLPCC and CMFCC with/without LDA. (T_d is the threshold used in the PCC algorithm; N_L and N_M are the total number of representative feature vectors for CLPCC and CMFCC, respectively)

	T_d	CLPCC	CLPCC +LDA	CMFCC	CMFCC +LDA	N_L	N_M
B420	0.5	29%	43%	54%	62%	504	561
	0.25	34%	46%	64%	70%	677	966
	0.1	43%	53%	79%	82%	1249	2334
	0.05	51%	59%	84%	87%	2279	4543
B561	0.5	32%	49%	60%	68%	609	681
	0.25	35%	51%	67%	74%	796	1035
	0.1	44%	55%	78%	82%	1372	2329
	0.05	51%	60%	84%	86%	2356	4536

Table 4
The reduced feature dimension after applying LDA.

	T_d	CLPCC	CLPCC +LDA	CMFCC	CMFCC +LDA
B420	0.5	15	12	15	12
	0.25	15	11	15	12
	0.1	15	11	15	12
	0.05	15	11	15	12
B561	0.5	15	12	15	11
	0.25	15	11	15	12
	0.1	15	11	15	12
	0.05	15	11	15	12

4. Conclusion

In this paper we have proposed a method capable of identifying bird from the sounds they generate. Each syllable corresponding to a piece of vocalization is first segmented. Each segmented syllable is then divided into a set of overlapped frames. For each frame, the LPCCs/MFCCs are computed as the feature vector of this frame. The averaged LPCCs (ALPCC) and averaged MFCCs (AMFCC) over all frames within a syllable are calculated as the representative feature vector of the syllable. To model the variant characteristics of syllables segmented from the same bird song, a codebook consisting of several feature vectors are employed to identify the same bird species. LDA is further exploited to improve the classification accuracy at a lower vector dimension.

From the experimental results, we can see that AMFCC outperforms ALPCC. This result is consistent with that for speech recognition. If a codebook consisting of several representative feature vectors is used to model the syllables of the same bird species, the best classification accuracy is 87% for the identification of 420 bird species. We have carefully examined the syllables that were inaccurately classified. The main reasons for the classification errors are: (1) syllables are not well segmented; (2) the background noise will somewhat affect the feature extraction results. To solve the first problem, a more efficient and robust syllable segmentation and classification techniques are required in the future work. To deal with background noise, an enhanced feature extraction approach which is robust to noise is required.

Acknowledgment

This research was supported in part by the National Science Council of R.O.C. under contract NSC-94-2213-E-216-022 and Chung Hua University under contract CHU-93-TR-004.

References

[1] E. A. Brenowitz, D. Margoliash, and K. M. Nordeen, “An introduction to birdsong and the avian song system”, *Journal of Neurobiology*, Vol. 33, Issue 5, pp. 495-500, Nov. 1997.

[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, Aug. 1996.

[3] J. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study", *Journal of the Acoustical Society of America*, Vol. 103, No. 4, pp. 2187-2196, Apr. 1998.

[4] A. L. McIlraith and H. C. Card, "Birdsong recognition with DSP and neural networks", in *Proceedings of IEEE Conference on Communications, Power, and Computing*, Vol. 2, pp. 409-414, May 1995.

[5] A. L. McIlraith and H. C. Card, "A comparison of backpropagation and dtatistical classifiers for bird identification", in *Proceedings of IEEE International Conference on Neural Networks* , Vol. 1, pp. 100-104, June 1997.

[6] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics", *IEEE Trans. on Signal Processing*, Vol. 45, No. 11, pp. 2740-2748, Nov. 1997.

[7] A. L. McIlraith and H. C. Card, "Bird song identification using artificial neural networks and statistical analysis", in *Proceedings of Canadian Conference on Electrical and Computer Engineering*, Vol. 1, pp. 63-66, May 1997.

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

[9] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[10] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized Mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 7, No. 5, pp. 525-532, Sep. 1999.

[11] J. W. Picone, "Signal modeling techniques in speech recognition", *Proceedings of the IEEE*, Vol. 81, pp. 1215–1247, 1993.

[12] N. Akrout, C. Diab, R. Prost, and R. Goutte, "A fast algorithm for vector quantization: application to codebook generation in image subband coding", *Signal Processing VI: Theories and Application*, Vol. 3, pp. 1227-1230, 1992.

[13] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York:Wiley, 2000.

[14] A. J. Doupe and P. K. Kuhl, "Birdsong and human speech: common themes and mechanisms", *Annual Review of Neuroscience*, Vol. 22, pp. 567-631, 1999.