Automatic Classification of Bird Species From Their Sounds Using Two-Dimensional Cepstral Coefficients

Chang-Hsing Lee, Chin-Chuan Han, *Member, IEEE*, and Ching-Chien Chuang

***Abstract*—This paper presents a method for automatic classification of birds into different species based on the audio recordings of their sounds. Each individual syllable segmented from continuous recordings is regarded as the basic recognition unit. To represent the temporal variations as well as sharp transitions within a syllable, a feature set derived from static and dynamic two-dimensional Mel-frequency cepstral coefficients are calculated for the classification of each syllable. Since a bird might generate several types of sounds with variant characteristics, a number of representative prototype vectors are used to model different syllables of identical bird species. For each bird species, a model selection method is developed to determine the optimal mode between Gaussian mixture models (GMM) and vector quantization (VQ) when the amount of training data is different for each species. In addition, a component number selection algorithm is employed to find the most appropriate number of components of GMM or the cluster number of VQ for each species. The mean vectors of GMM or the cluster centroids of VQ will form the prototype vectors of a certain bird species. In the experiments, the best classification accuracy is 84.06% for the classification of 28 bird species.**

***Index Terms*—Birdsong classification, Gaussian mixture models (GMMs), two-dimensional Mel-frequency cepstral coefficients.**

I. Introduction

IN DAILY life, we can hear a variety of creatures' sounds, including human speech, dog barks, birdsong, cicada sounds, frog calls, 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. Classification of animals by their sounds is valuable for biological research and environmental monitoring applications, especially in detecting and locating animals. A common way for the biologists to assess the environmental impact of human activities on animals is to detect, locate, identify, and count animals in a site. Since birds can be observed easily by experienced bird watchers, many biologists intend to identify and count birds in a specific area to estimate long-term population trends of different bird species. However, most people often hear the sounds made by birds rather than see the birds themselves. In fact, most of the bird vocalizations have evolved to be species-specific. Therefore, it is a natural and adequate way to automatically identify bird species from their vocalizations.

In general, most bird vocalizations are short and unmusical and thus cannot be termed as "songs."[1] These sounds are generally referred to as "calls." These calls have considerable functionalities, including alarm calls, distress calls, aggressive calls, territorial calls, flight calls, etc. In general, the sounds that birds make and their syntactical arrangements change significantly from bird to bird. Therefore, birdsong is typically represented by a hierarchical acoustic structure [1]. The different structural components of birdsong 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 durations is a song *bout*.

Most of the traditional studies of bird sounds are based on visual inspection of sound spectrograms (sonograms). To continuously identify the spectrograms of a large set of bird sounds by human experts is an extremely laborious and time-consuming task. Thus, automatic recognition of bird sounds might be desirable. Anderson *et al.* [2] used dynamic time warping (DTW) for automated analysis of continuous recordings of birdsong. They directly compared signal spectrograms, and identified constituents and constituent boundaries. The feature vectors were derived from log magnitudes of fast Fourier transform (FFT) bins from 0.5 to 10 KHz. They evaluated the performance on the vocalizations of indigo buntings (*Passerina cyanea*) and zebra finches (*Taeniopygia guttata*). The test data was collected from a low-cutter, low-noise environment. The representative templates (syllables) were manually labeled by human experts. They identified syllables in stereotyped songs and calls with greater than 97% accuracy. Syllables in the more variable and lower amplitude plastic songs were identified with approximate 84% accuracy.

Kogan and Margoliash [3] compared two techniques, DTW and hidden Markov models (HMMs), for automatic recognition of birdsong elements from continuous recordings. The feature vectors used for DTW classification were the log magnitudes of FFT bins from 0.5 to 10 KHz. Six types of

Manuscript received February 20, 2008; revised July 19, 2008. Current version published October 17, 2008. This work was supported in part by the National Science Council of R.O.C. under Contract NSC-96-2221-E-216-043. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. George Tzanetakis.

C.-H. Lee and C.-C. Chuang are with the Department of Computer Science and Information Engineering, Chung Hua University, Hsinchu 300, Taiwan (e-mail: chlee@chu.edu.tw; m09402051@chu.edu.tw).

C.-C. Han is with the Department of Computer Science and Information Engineering, National United University, Miao-Li 360, Taiwan (e-mail: cchan@nuu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2005345

[1][Online]. Available: http://www.earthlife.net/birds/song.html.

feature parameters were compared on HMMs performance, including linear predictive coding (LPC) coefficients, LPC cepstral coefficients (LPCCs), LPC reflection coefficients, Mel-frequency cepstral coefficients (MFCCs), log mel-filter bank channels, and linear mel-filter bank channels. Among these feature sets, the best results were achieved for MFCC. Therefore, the MFCC parameters, including energy, first, and second derivatives, were used for all experiments. Experiment results showed 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. In most experiments, HMMs exhibited better performance than DTW. One disadvantage of HMMs 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 different birds (song sparrow, fox sparrow, marsh wren, sedge wren, yellow warbler, and red-winged blackbird). Temporal parameters as well as spectral information were used as features in their study. The temporal parameters include the number of elements, the mean and standard deviation of element lengths, and the mean and standard deviation of silence lengths within each song. The spectral information includes LPC coefficients [4] or means and standard deviations of the power spectral density for nine spectral bands [5]–[7]. Quadratic discriminant analysis was exploited to boost the classification accuracy. The classification accuracy was 82% using back-propagation neural network and 93% using quadratic discriminant analysis.

Härmä proposed a method for automatic identification of bird species based on sinusoidal modeling of syllables [8]. Each syllable was approximated as amplitude and frequency trajectories. A segmentation algorithm was proposed to divide each sound into a number of syllables. The weighted sum of differences of frequency and amplitude trajectories was used as the distance criterion between two syllables. Experimental results showed that with a limited number of bird species, a recognizer based on this signal model may be sufficient.

Since many bird sounds with a clear harmonic spectrum structure cannot be well modeled by pure sinusoids, Härmä and Somervuo proposed a method to classify bird syllables into four classes according to their harmonic structure [9]. In their experimental results, the recognition accuracy could be improved by 5%–20% for many bird species. However, the improvements were limited for some species. Therefore, they conceived that a better recognition accuracy could be achieved based on the song-level structure instead of isolated syllables only. Thus, they proposed a birdsong recognition approach based on syllable pair histograms [10]. A set of Gaussian syllable prototypes would be automatically found to represent each syllable. The syllable pair histogram, which reveals some temporal structure of each birdsong, was collected from the Gaussian syllable prototypes of every two consecutive syllables (called syllable pairs). Each birdsong was then modeled by the syllable-pair histogram. Finally, the mutual correlation between two histograms was employed to compare the similarity of two histograms. In their experiments, 257 songs derived from 50 bird individuals belonging to four species (*Fricoe*, *Phylus*, *Phycol*, and *Parmaj*) were used for performance evaluation. The classification accuracy was 76%, 79%, and 80% for three different Gaussian syllable prototype sets containing 10, 30, and 50 Gaussians, respectively.

Recently, Somervuo *et al.* compared three feature representations of bird sounds for automatic species recognition: sinusoidal model, MFCC model, and descriptive parameters [11]. The best result for single-syllable-based recognition was obtained by using trajectory model with MFCC vectors. They also shown that the recognition results could be significantly improved using song-based recognition, i.e., the classification is based on a sequence of consecutive syllables instead of a single syllable.

Singh and Theunissen analyzed natural sounds by calculating the probability distributions of the amplitude envelope of the sounds and their time–frequency correlations given by the modulation spectra [12]. The modulation spectra were obtained by calculating the two-dimensional Fourier transform of the autocorrelation matrix of their spectrograms. It was shown that most of the spectral modulation power concentrates on those components with slow temporal modulation for animal vocalizations and human speech. Wang *et al.* analyzed birdsongs using the spectral correlation between the power at two frequencies with a time lag $t, C_s(f, f', t)$ [13]. A special case of the spectral correlation appeared in the neuroscience is the instantaneous correlations, $C_s(f, f', 0)$. The so-called "modulation spectrum" can be obtained by computing the average of $C_s(f + f', f', t)$ over frequencies f' followed a two-dimensional Fourier transform. This is equal to the average of the modulus squared 2-D cepstrum. $C_s(f, f, t)$, another special case which computes the correlation function of spectral power at a specific frequency, provides one way to identify rhythmic information in birdsongs.

In this paper, both static and dynamic two-dimensional MFCC (TDMFCC) are calculated as vocalization features for the classification of bird species from their songs or calls. These features exploit human auditory perception on different frequency tones as well as the temporal evolution of the analyzed sounds. To model different syllables generated by the same bird species, GMM or VQ is used to find a number of representative prototype vectors for each bird species. A model selection method is developed to determine the optimal mode between GMM and VQ when the amount of training data is different for each species. In addition, a component number selection algorithm is employed to find the most appropriate number of components of GMM or the cluster number of VQ for each species. In the following section, we will describe the proposed bird sound classification method. Experimental results will be presented in Section III to show the effectiveness of the proposed method. Finally, a conclusion is given in Section IV.

II. Proposed Bird Sound Classification System

In this paper, we focus on automatic classification of individual syllables from birdsong recordings. The syllables are manually labeled from each birdsong recording. Each syllable is regarded as the elementary unit for the classification of bird species. The classification system (see Fig. 1) consists of two

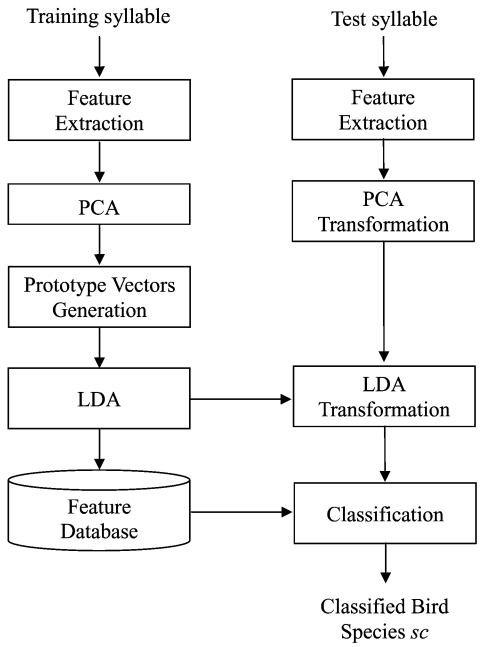


Fig. 1. Block diagram of the proposed birdsong classification system.

phases: the training phase and the classification phase. The training phase is composed of four main modules: feature extraction, principal component analysis (PCA), prototype vectors generation, and linear discriminant analysis (LDA). The classification phase consists of three modules: feature extraction, PCA transformation, LDA transformation, and classification. A detailed description of each module will be described below.

A. Feature Extraction

For most bird sounds, within each syllable there exist more or less temporal variations among neighboring analyzed frames. To describe the temporal variations, TDMFCC is employed to describe both static and dynamic characteristics of a syllable. In addition, it has been shown that the sound part with maximum spectral transition contains most important information for syllables recognition. Thus, dynamic TDMFCC (DTDMFCC) will be developed to describe sharp transitions within a syllable.

1) TDMFCC: Two-dimensional cepstrum (TDC) was originally used for automatic recognition of speech signals due to its ability to implicitly represent both static feature (instantaneous cepstrum) and dynamic feature (temporal variations) of a speech signal in a matrix form [14]–[16]. The TDC matrix $T(q,n)$ can be obtained by applying 2-D discrete cosine transform (DCT) to a sequence of consecutive logarithmic spectrum. The first dimension q of the TDC matrix represents cepstrum (quefrency), and the second one n captures the temporal variations of each cepstral coefficient. In this paper, TDMFCC is employed to model each syllable of bird sounds. Instead of applying 2-D DCT on the logarithmic spectrum, we apply 2-D DCT on the logarithmic energies of *Mel*-scale bandpass filters, defined according to a model of human auditory perception, to obtain the TDMFCC matrix $C(q,n)$:

$$C(q,n) = \sum_{t=0}^{L-1}\sum_{b=0}^{B-1} \log(E_t(b)) \times \cos\left(\frac{(2b+1)q\pi}{2B}\right)\cos\left(\frac{(2t+1)n\pi}{2L}\right) \quad 0 \le q < B, 0 \le n < L \quad (1)$$

where $E_t(b)$ is the energy of the bth *Mel*-scale bandpass filter of the tth frame, q is the quefrency index, n is the index of modulation frequency, B is the number of *Mel*-scale bandpass filters, and L is the number of frames within a syllable. Since 2-D DCT can be decomposed into two 1-D DCTs, $C(q,n)$ can also be obtained by applying 1-D DCT to a sequence of L consecutive MFCC coefficient along time axis

$$C(q,n) = \sum_{t=0}^{L-1} c_t(q)\cos\left(\frac{(2t+1)n\pi}{2L}\right) \quad 0 \le q < B, 0 \le n < L \quad (2)$$

where $c_t(q)$ is the qth MFCC coefficient of the tth frame

$$c_t(q) = \sum_{b=0}^{B-1} \log(E_t(b))\cos\left(\frac{(2b+1)q\pi}{2B}\right). \quad (3)$$

The first row of the TDMFCC matrix with index $q = 0$ preserves temporal variations of short-time energy. Each element in the first column with index $n = 0$ represents the average values of each cepstral coefficient of all analyzed frames. For human speech signal, along the quefrency axis q, the lower coefficients represent the spectral envelope and the higher ones represent the pitch and excitation. Along the time axis n, the lower coefficients represent the global variation of the quefrencies, whereas the higher ones represent the local variation of the quefrencies.

Since the durations are not identical for different syllables, the number of analyzed frames differs from syllable to syllable. Therefore, the number of columns in $C(q,n)$ differs for distinct syllables. In fact, the coefficients at lower part along the quefrency axis q and time axis n provide more useful information for sound recognition than those coefficients at higher part. Thus, we take the coefficients from the first 15 rows and the first five columns of $C(q,n)$ excluding the dc coefficient $C(0,0)$ as the preliminary sound features of a syllable. In total, 74 coefficients were selected from the TDMFCC matrix $C(q,n)$ to form the TDMFCC feature vector of a syllable. Thus, the dimension of the feature vector is fixed regardless of the duration of a syllable. In summary, the TDMFCC feature vector can be represented as

$$\mathbf{F}_{\mathrm{TDMFCC}} = [C(0,1),\ldots,C(0,4),C(1,0),\ldots, C(1,4),\ldots,C(14,0),\ldots,C(14,4)]^{\mathrm{T}}. \quad (4)$$

2) DTDMFCC: In addition to the features derived from TDMFCC, DTDMFCC is used to emphasize sharp transitions within a syllable. DTDMFCC is motivated by the prior work

proposed by Furui [17] in which a speaker-independent isolated word recognition method was developed based on the combination of instantaneous and dynamic features of speech spectrum. Their experiment on isolated syllable recognition demonstrated that the portion of the utterance with maximum spectral transition bears the most important phonetic information in all syllables. Therefore, the dynamic features derived from the regression coefficients, defined as the first-order orthogonal polynomial coefficients, were employed for isolated word recognition. These regression coefficients represent the slope of the time function of each cepstral coefficient within the speech segment being measured.

In this paper, DTDMFCC is extracted to highlight the portion of maximum spectral transitions within a syllable. The bth regression coefficient of the tth frame is first computed as

$$r_t(b) = \frac{\sum_{i=1}^{n_0} i \cdot |E_{t+i}(b) - E_{t-i}(b)|}{\sum_{i=-n_0}^{n_0} i^2}, \quad 0 \le b < B \tag{5}$$

where n_0 is the interval length for measuring the transitional information. From the above definition, we can see that $r_t(b)$ represents the energy transition around the tth frame, for the bth *Mel*-scale bandpass filter output. To emphasize the portion of the sound with maximum spectral transitions, each regression coefficient, $r_t(b)$, is added to $E_t(b)$, to obtain the enhanced energy, $\hat{E}_t(b)$

$$\hat{E}_t(b) = E_t(b) + r_t(b), \quad 0 \le b < B. \tag{6}$$

DTDMFCC is then derived by applying 2-D DCT to the logarithmic emphasized energy, $\log[\hat{E}_t(b)]$

$$\hat{C}(q,n) = \sum_{t=0}^{L-1}\sum_{b=0}^{B-1} \log(\hat{E}_t(b)) \cos\left(\frac{(2b+1)q\pi}{2B}\right)\cos\left(\frac{(2t+1)n\pi}{2L}\right). \tag{7}$$

Similarly, the coefficients selected from the first 15 rows and the first five columns of $\hat{C}(q,n)$ excluding the dc coefficient C (0,0) are taken as the DTDMFCC features of a syllable. Therefore, the DTDMFCC feature vector can be represented as

$$\mathbf{F}_{\mathrm{DTDMFCC}} = [\hat{C}(0,1),\ldots,\hat{C}(0,4),\hat{C}(1,0),\ldots, \hat{C}(1,4),\ldots,\hat{C}(14,0),\ldots,\hat{C}(14,4)]^{\mathrm{T}}. \tag{8}$$

Note that TDMFCC is obtained by applying 2-D DCT to logarithmic energy $\log[E_t(b)]$.

3) Combination of Feature Vectors: To achieve better classification results, we combine the above two described feature vectors ($\mathbf{F}_{\mathrm{TDMFCC}}$ and $\mathbf{F}_{\mathrm{DTDMFCC}}$) together to obtain a larger feature vectors, notated $\mathbf{F}_{\mathrm{SDTDMFCC}}$. The combined feature vector $\mathbf{F}_{\mathrm{SDTDMFCC}}$ can describe the static, dynamic, and spectral transitional information within a syllable. In this paper, $\mathbf{F}_{\mathrm{SDTDMFCC}}$ is formed by concatenating $\mathbf{F}_{\mathrm{TDMFCC}}$ and $\mathbf{F}_{\mathrm{DTDMFCC}}$ together, that is

$$\mathbf{F}_{\mathrm{SDTDMFCC}} = \left[\mathbf{F}_{\mathrm{TDMFCC}}^{\mathrm{T}}, \mathbf{F}_{\mathrm{DTDMFCC}}^{\mathrm{T}}\right]^{\mathrm{T}}. \tag{9}$$

4) Normalization of Feature Values: Without loss of generality, let $\mathbf{F}$ denote the calculated feature vector ($\mathbf{F}_{\mathrm{TDMFCC}}$ or $\mathbf{F}_{\mathrm{DTDMFCC}}$ or $\mathbf{F}_{\mathrm{SDTDMFCC}}$) of a syllable. Since the dynamic range of each feature value is not identical, each feature value will be normalized to make the range of each normalized value between 0 and 1

$$x(m) = \begin{cases} 0, & F(m) \le Q_1(m) \\ \frac{F(m)-Q_1(m)}{Q_3(m)-Q_1(m)}, & Q_1(m) < F(m) < Q_3(m) \\ 1, & F(m) \ge Q_3(m) \end{cases} \tag{10}$$

where $F(m)$ is the mth feature value, $x(m)$ is the normalized mth feature value, $Q_1(m)$ (or $Q_3(m)$) denotes the first (or third) quartile which is defined as the value such that 25 (or 75) percent of the mth feature values of all training syllables are less than or equal to it. According to (10), those extremely high and extremely low feature values will be normalized to be 1 and 0 such that the normalized feature values are less affected by noisy sound. The first quartile $Q_1(m)$ and the third quartile $Q_3(m)$ are computed for each feature value and these values are stored for later reference. In the classification phase, for actual normalization, each feature value extracted from the input syllable is modified using the reference quartile values ($Q_1(m)$ and $Q_3(m)$) to obtain normalized values using (10).

B. Principal Component Analysis (PCA)

PCA has been a widely used technique for dimensionality reduction [18]. PCA is defined as the orthogonal projection of the data onto a lower dimensional vector space such that the variance of the projected data is maximized. First, the D-dimensional mean vector $\boldsymbol{\mu}$ and D × D covariance matrix $\boldsymbol{\Sigma}$ are computed for the set of D-dimensional training vectors $X = \{\mathbf{x}_j, j = 1,\ldots,N\}$

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j \tag{11}$$

$$\sum = \frac{1}{N}\sum_{j=1}^{N}(\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})^{\mathrm{T}}. \tag{12}$$

Second, the eigenvectors and corresponding eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ are computed and sorted in decreasing order of eigenvalues. Let the eigenvector $\mathbf{v}_i$ be associated with eigenvalue $\lambda_i, 1 \le i \le \mathrm{D}$. The first d eigenvectors having the largest eigenvalues are the columns of the D × d transformation matrix $\boldsymbol{A}_{\mathrm{PCA}}$

$$\boldsymbol{A}_{\mathrm{PCA}} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d]. \tag{13}$$

The number of selected eigenvectors d can be determined by finding the minimum integer that satisfies the following criterion:

$$\sum_{j=1}^{d}\lambda_j \ge \alpha\sum_{j=1}^{D}\lambda_j \tag{14}$$

where α determine how many percentage of information need to be preserved. The projected vector can be computed according to the transformation matrix $\boldsymbol{A}_{\mathrm{PCA}}$

$$\mathbf{x}_{\mathrm{PCA}} = \boldsymbol{A}_{\mathrm{PCA}}^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\mu}). \tag{15}$$

C. Prototype Vectors Generation

As mentioned previously, the sound of each bird consists of a number of syllables having distinct characteristics. That is, any two syllables segmented from identical bird sound might differ significantly. Therefore, the feature vectors extracted from the syllables of identical bird species will reveal many isolated manifolds in the feature space. As a result, modeling the sound of each bird species with a single feature vector is bound to fail. A better approach to cope with this problem is to model the sounds of each species with a number of representative prototype vectors. These prototype vectors can be obtained by classifying all syllables derived from identical bird species into some subcategories such that syllables with similar feature vectors are clustered together. Vector quantization (VQ) and Gaussian mixture model (GMM) have been widely used as the speaker model in speaker recognition systems [19]. In general, GMM usually achieves higher recognition accuracy than VQ when the available training data is sufficient. However, GMM becomes ineffective if the amount of training data is insufficient to reliably estimate the covariance matrices of mixture densities. In practice, it has been shown that VQ outperforms GMM when the obtainable training data is limited [20]. Since the training data size differs from species to species in our birdsong database (see Table I). That is, some bird species will achieve higher classification accuracy when GMM is used as the acoustic model, whereas VQ-based models will perform better for some other bird species. To choose the best acoustic model (GMM or VQ) for each bird species, we will employ the statistical speaker model selection (SMSS) method proposed by Nishida and Kawahara [21]. The selection mechanism is based on the Bayesian information criterion (BIC) evaluated for GMM and VQ-based models. Another problem must be considered is how many prototype vectors are sufficient to describe the acoustic variations of each bird species. Cheng *et al.* have proposed a self-splitting Gaussian mixture learning (SGML) algorithm to find an appropriate number of components of GMM based on a self-splitting validity measure described in terms of BIC [22]. In this paper, the SGML algorithm is employed to determine the appropriate number of components of GMM or the cluster number of VQ. To simplify the classification task, the mean vectors of all Gaussian components of GMM or the cluster centroids of VQ will form the prototype vectors of a bird species.

1) Acoustic Model Selection Method: The SMSS method was proposed to automatically select the optimal model (GMM or VQ) based on BIC, for different training data size. The SMSS method will select a discrete VQ model when the training data is sparse and will seamlessly switch to the continuous GMM model when the obtained training data is sufficient. Since the model structure and distance measure are different for GMM and VQ, the extended VQ (EVQ) [23], which assigns the same mixture weight and covariance to all Gaussian mixture components, was used to cope with the structure inconsistencies between GMM and VQ. Therefore, the distance measure of VQ can be compared to the likelihood of GMM. In fact, EVQ becomes VQ if the covariance matrix is replaced with the identity matrix.

TABLE I
COMMON AND LATIN NAMES OF BIRD SPECIES OF THE BIRDSONG DATABASE AND THEIR CORRESPONDING SYLLABLE NUMBERS IN THE TRAINING SET AND TEST SET

Common Name	Latin Name	Training Syllables	Test Syllables
Crested Serpent Eagle	Spilornis cheela	10	4
Bronzed Drongo	Dicrurus aeneus	229	37
Gray-headed Pygmy Woodpecker	Dendrocopos canicapillus	17	25
Blue Shortwing	Brachypteryx montana	296	29
Streak-breasted Scimitar Babbler	Pomatorhinus ruficollis	120	22
Taiwan Firecrest	Regulus goodfellowi	194	57
Taiwan Sibia	Heterophasia auricularis	98	14
White-throated Laughing Thrush	Garrulax albogularis	100	37
White-breasted Water Hen	Amaurornis phoenicurus	172	15
Beavan's Bullfinch	Pyrrhula erythaca	70	8
Gray-sided Laughing Thrush	Garrulax caerulatus	31	31
Alpine Accentor	Prunella collaris	122	53
Green-backed Tit	Parus monticolus	140	14
Taiwan Yuhina	Yuhina brunneiceps	49	12
Red-headed Tit	Aegithalos concinnus	61	24
Collared Bush Robin	Erithacus johnstoniae	230	18
Taiwan Bulbul	Pycnonotus taivanus Styan	131	30
Taiwan Hill Partridge	Arborophila crudigularis	123	27
Verreaux's Bush Warbler	Cettia acanthizoides	51	8
Oriental Cuckoo	Cuculus saturatus	284	45
Taiwan Tit	Parus holsti	222	27
Vivid Niltava	Niltava vivida	76	12
Coal Tit	Parus ater	149	34
Crested Goshawk	Accipiter trivirgatus	32	16
Gould's Fulvetta	Alcippe brunnea	32	18
Collared Pigmy Owlet	Glaucidium brodiei	61	14
Swinhoe's Pheasant	Lophura swinhoii	23	10
Steere's Liocichla	Liocichla steerii	20	5
Total syllable number		3143	646

For GMM, only the diagonal elements of the covariance matrix are used. The BIC of the GMM model for bird species s is calculated as follows:

$$\mathrm{BIC}_{\mathrm{GMM}}^{(s)} = \log P\left(X \,\middle|\, \lambda_{\mathrm{GMM}}^{(s)}\right) - \frac{1}{2}M(2d+1)\log N \tag{16}$$

where $X = \{\mathbf{x}_j \,|\, 1 \le j \le N\}$ is the set of training vectors, $\lambda_{\mathrm{GMM}}^{(s)} = \{\omega_r^{(s)}, \boldsymbol{\mu}_r^{(s)}, \sum_r^{(s)} | 1 \le r \le M\}$ is the parameter set of GMM, $\log P(X \,|\, \lambda_{\mathrm{GMM}}^{(s)})$ denotes the log likelihood of training set X modeled by the GMM with parameter set $\lambda_{\mathrm{GMM}}^{(s)}$, M is the number of mixture components, d is the dimension of each feature vector, and N is the number of training vectors.

For EVQ, the mixture weights are identically assigned as $\omega_{\mathrm{EVQ}} = 1/M$. In addition, the covariance matrix of each Gaussian component is replaced with the average covariance matrix of GMMs of all bird species

$$\sum\nolimits_{\mathrm{EVQ}} = \frac{\sum_{s=1}^{S}\sum_{j=1}^{N_s}\sum_{\mathrm{GMM}_j}^{(s)}}{\sum_{s=1}^{S} N_s} \tag{17}$$

where S is the number of bird species, N_s is the number Gaussian components selected for bird species s, and $\sum_{\mathrm{GMM}_j}^{(s)}$ is the covariance matrix of the jth Gaussian component of bird species s. The BIC for the EVQ model is then calculated as follows:

$$\mathrm{BIC}_{\mathrm{EVQ}}^{(s)} = \log P\left(X \left| \lambda_{\mathrm{EVQ}}^{(s)} \right.\right) - \frac{1}{2}(M+1)d\log N \quad (18)$$

where d-dimenaional mean vectors of M Gaussian components and one common diagonal covariance matrix are counted. If $\mathrm{BIC}_{\mathrm{GMM}}^{(s)} > \mathrm{BIC}_{\mathrm{EVQ}}^{(s)}$, GMM will be selected as the best model for bird species s; otherwise, VQ will be the selected model. From the definitions of (16) and (18), the VQ-based model will be selected when the training data is limited because its model complexity is small. If a large number of training data is available, GMM is expected to be selected because its likelihood is larger.

2) Component Number Selection Method: GMM can be thought of as a probabilistic model-based clustering approach in which the training data is to be grouped into clusters by assigning each training sample to the Gaussian component which is most likely to generate it. The most widely used approach to estimate the parameters of GMM is the expectation maximization (EM) algorithm [19]. However, the EM algorithm assumes that the number of Gaussian components (cluster number) is known beforehand. In fact, some kind of birds can generate only a few regular sound patterns, whereas some bird species will generate a considerable number of different sounds. As a result, the cluster number used to model the sounds of different birds must differ from species to species. That is, it depends on the acoustic variations of each bird species to determine the cluster number. Traditionally, the cluster number is usually determined empirically. However, the selection of cluster number used to model the sound of each bird species will affect the classification accuracy. How to automatically determine the cluster number is still an important issue related to the EM algorithm. In this paper, we will apply the SGML algorithm [22] proposed by Cheng *et al.* to cope with this problem. The SGML algorithm starts with a single Gaussian component and successively splits one selected component into two new ones. The selection and split process is repeated until the most appropriate number of components is found. In SGML, BIC is employed to find the component to be split and to determine the appropriate number of components. A detailed description of the SGML algorithm can be found in [22].

D. Linear Discriminant Analysis (LDA)

To further improve the classification accuracy at a lower dimensional feature space, LDA [18] is employed to provide higher discriminability among various bird species (classes). LDA tries to minimize the within-class distance while maximizing the between-class distance. In LDA, an optimal transformation matrix corresponding to a mapping from a d-dimensional feature space to a k-dimensional space is determined, where $k \leq d$. The most widely used transformation matrix is a linear mapping that maximizes the so-called Fisher criterion J_F:

$$J_F(\boldsymbol{A}) = \mathrm{tr}((\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_W\boldsymbol{A})^{-1}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{A})) \quad (19)$$

where $\boldsymbol{S}_W$ and $\boldsymbol{S}_B$ are the within-class scatter matrix and between-class scatter matrix, respectively. The within-class scatter matrix is defined as

$$\boldsymbol{S}_W = \sum_{s=1}^{S}\sum_{\mathbf{x}\in C_s}(\mathbf{x}-\boldsymbol{\mu}_s)(\mathbf{x}-\boldsymbol{\mu}_s)^{\mathrm{T}} \quad (20)$$

where S is the total number of classes (bird species), C_s is the set of feature vectors assigned to class s, $\boldsymbol{\mu}_s$ is the mean vector of class s. The between-class scatter matrix is given by

$$\boldsymbol{S}_B = \sum_{s=1}^{S} N_s(\boldsymbol{\mu}_s-\boldsymbol{\mu})(\boldsymbol{\mu}_s-\boldsymbol{\mu})^{\mathrm{T}} \quad (21)$$

where N_s is the number of feature vectors in class s, $\boldsymbol{\mu}$ is the mean vector of all training vectors. From (19), we can see that LDA tries to find a transformation matrix that maximizes the ratio of between-class scatter to within-class scatter. In this paper, a whitening procedure is combined with LDA to transform the multivariate normal distribution of the set of training vectors into a spherical one [18]. First, the eigenvectors and corresponding eigenvalues of $\boldsymbol{S}_W$ are calculated. Let $\boldsymbol{\Phi}$ denote the matrix whose columns are the orthonormal eigenvectors of $\boldsymbol{S}_W$, and $\boldsymbol{\Lambda}$ the diagonal matrix of the corresponding eigenvalues. Thus, $\boldsymbol{S}_W\boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{\Lambda}$. Each training vector $\mathbf{x}$ is then whitening transformed by $\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}$

$$\mathbf{x}' = (\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2})^{\mathrm{T}}\mathbf{x}. \quad (22)$$

Thus, the within-class scatter matrix $\boldsymbol{S}'_W$ of the whitened vectors will become an identity matrix since

$$\begin{aligned}\boldsymbol{S}'_W &= \sum_{s=1}^{S}\sum_{\mathbf{x}\in C_s}(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2})^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu}_s)(\mathbf{x}-\boldsymbol{\mu}_s)^{\mathrm{T}}(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}) \\ &= (\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2})^{\mathrm{T}}\boldsymbol{S}_W(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}) \\ &= (\boldsymbol{\Lambda}^{-1/2})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{S}_W\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2} \\ &= \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} \\ &= \boldsymbol{\Lambda}^{-1/2}\mathbf{I}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} \\ &= \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} \\ &= \mathbf{I}. \end{aligned} \quad (23)$$

Thus, the whitened between-class scatter matrix $\boldsymbol{S}'_B = (\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2})^{\mathrm{T}}\boldsymbol{S}_B(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2})$ contains all the discriminative information. A transformation matrix $\boldsymbol{\Psi}$ can be determined by finding the eigenvectors of $\boldsymbol{S}'_B$. Assuming that the eigenvalues are sorted in a decreasing order, the eigenvectors corresponding to the largest $k = (S-1)$ eigenvalues will constitute the

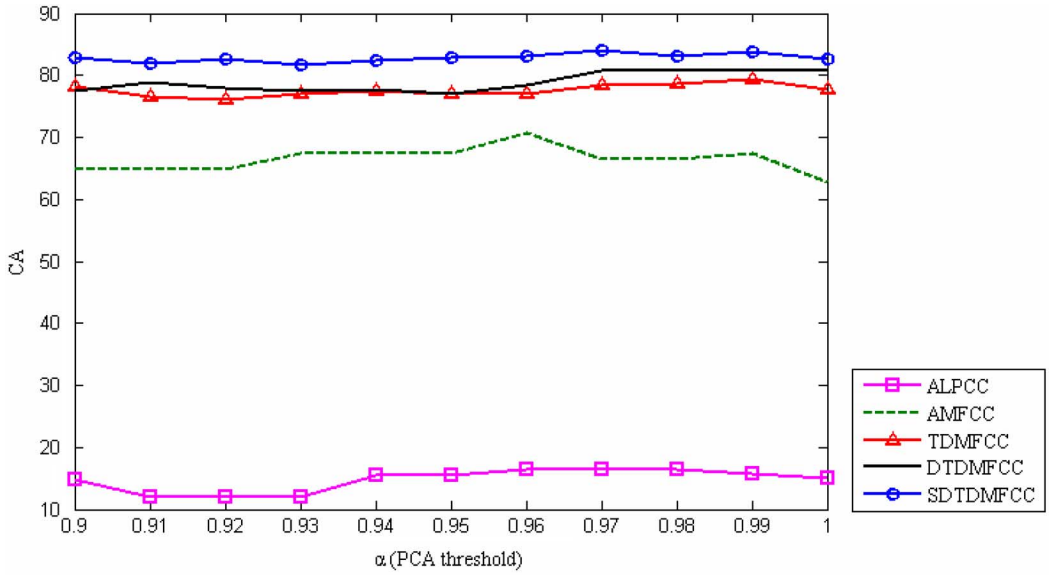


Fig. 2. Comparison of classification results for different PCA threshold α.

columns of transformation matrix $\mathbf{\Psi}$. Finally, the optimal transformation matrix $\boldsymbol{A}_{\mathrm{LDA}}$ is defined as

$$\boldsymbol{A}_{\mathrm{LDA}} = \mathbf{\Phi}\mathbf{\Lambda}^{-1/2}\mathbf{\Psi}. \tag{24}$$

$\boldsymbol{A}_{\mathrm{LDA}}$ is employed to transform each PCA transformed d-dimensional feature vector to be a lower k-dimensional vector. Let $\mathbf{x}_{\mathrm{PCA}}$ denote a d-dimensional PCA transformed vector, the LDA transformed k-dimensional feature vector can be computed by

$$\mathbf{f} = \boldsymbol{A}_{\mathrm{LDA}}^{\mathrm{T}}\mathbf{x}_{\mathrm{PCA}}. \tag{25}$$

E. Classification Phase

In this paper, the classification of each individual syllable is based on the nearest neighbor classifier. In the classification phase, the feature vector of each input syllable is first computed. Identical normalization using (10) is applied to each feature value. The normalized feature vector $\mathbf{x}$ is then transformed by using the PCA transformation matrix $\boldsymbol{A}_{\mathrm{PCA}}$ followed by the LDA transformation matrix $\boldsymbol{A}_{\mathrm{LDA}}$ to obtain the final feature vector $\mathbf{f}$

$$\mathbf{f} = \boldsymbol{A}_{\mathrm{LDA}}^{\mathrm{T}}\boldsymbol{A}_{\mathrm{PCA}}^{\mathrm{T}}\mathbf{x}. \tag{26}$$

The distance between $\mathbf{f}$ and every prototype vector of every bird species is measured by the Euclidean distance. The subject code sc that represents the classified bird species is determined by finding the prototype vector that has the minimum distance to $\mathbf{f}$

$$sc = \arg\min_{s} d(\mathbf{f}, \mathbf{f}_{s,j}), \quad 1 \le s \le S, \quad 1 \le j \le N_s \tag{27}$$

where $\mathbf{f}_{s,j}$ is the jth prototype vector of the sth bird species, and N_s is the number of prototype vectors for the sth bird species.

III. Experimental Results

The birdsong database, which contains the birdsong of 28 bird species in Taiwan, collected from commercially available compact disks and the Internet are used for the experiments [24], [25].[2] The sounds are field recordings with additional background sounds/noise. Some recordings contain vocalizations generated by multiple individuals. The sampling frequency is 44 100 Hz with each sample digitized in 16-bit accuracy. To test whether the proposed approach is individually independent for bird species classification, the test data and training data obtained from different recordings are used in this experiment. The syllables segmented from the compact disks [24], [25] are used as the training set, whereas the syllables derived from the website of National Fonghuanggu Bird Park[2] are used as the test set. In total, the training set contains 3143 syllables whereas the test set consists of 646 syllables. A summary of the recordings about these 28 bird species is shown in Table I. The syllables extracted from identical track are attributed to the same bird species. The performance is measured in terms of the classification accuracy (CA) defined as

$$\mathrm{CA} = \frac{N_{\mathrm{CA}}}{N_T} \times 100\% \tag{28}$$

where N_{CA} is the number of correctly classified syllables, and N_T is the total number of test syllables.

The experiments are implemented in Borland C++ Builder 6.0. The test platform is Intel Pentium-M, 1.6-GHz CPU, 1-G RAM with Windows XP Professional operating system. Several methods, including ALPCC, AMFCC [26], STDMFCC, DTDMFCC, and SDTDMFCC are conducted to compare their performance. For ALPCC or AMFCC, each syllable is represented by only one feature vector derived by averaging the calculated LPCC or MFCC of all frames in each syllable. Fig. 2 compares the classification results for different PCA threshold α defined in (14). From this figure, we can see that TDMFCC and DTDMFCC have comparable performance in terms of classification accuracy and they always outperform AMFCC and ALPCC. Further, SDTDMFCC, which combines TDMFCC

[2][Online]. Available: http://www.fhk.gov.tw.

TABLE II
SUMMARIZATION OF CLASSIFICATION ACCURACY (CA), SELECTED MODEL (EVQ OR GMM), THE CLUSTER NUMBER (N_S) FOR EACH BIRD SPECIES USING SDTDMFCC WHEN PCA THRESHOLD $\alpha = 0.97$

Subject Code	Bird Name	*CA* (%)	N_s	Selected Model
1	Crested Serpent Eagle	100.00	2	EVQ
2	Bronzed Drongo	86.49	5	EVQ
3	Gray-headed Pygmy Woodpecker	0.00	1	EVQ
4	Blue Shortwing	72.41	4	EVQ
5	Streak-breasted Scimitar Babbler	54.55	3	GMM
6	Taiwan Firecrest	100.00	3	EVQ
7	Taiwan Sibia	100.00	6	EVQ
8	White-throated Laughing Thrush	94.59	3	EVQ
9	White-breasted Water Hen	100.00	4	EVQ
10	Beavan's Bullfinch	100.00	3	EVQ
11	Gray-sided Laughing Thrush	100.00	3	EVQ
12	Alpine Accentor	71.70	1	EVQ
13	Green-backed Tit	7.14	5	EVQ
14	Taiwan Yuhina	100.00	3	EVQ
15	Red-headed Tit	100.00	2	EVQ
16	Collared Bush Robin	94.44	9	EVQ
17	Taiwan Bulbul	83.33	5	EVQ
18	Taiwan Hill Partridge	88.89	6	EVQ
19	Verreaux's Bush Warbler	100.00	4	EVQ
20	Oriental Cuckoo	95.56	3	GMM
21	Taiwan Tit	96.30	7	EVQ
22	Vivid Niltava	100.00	5	EVQ
23	Coal Tit	100.00	4	EVQ
24	Crested Goshawk	100.00	3	EVQ
25	Gould's Fulvetta	33.33	1	EVQ
26	Collared Pigmy Owlet	100.00	1	EVQ
27	Swinhoe's Pheasant	100.00	3	EVQ
28	Steere's Liocichla	80.00	3	EVQ

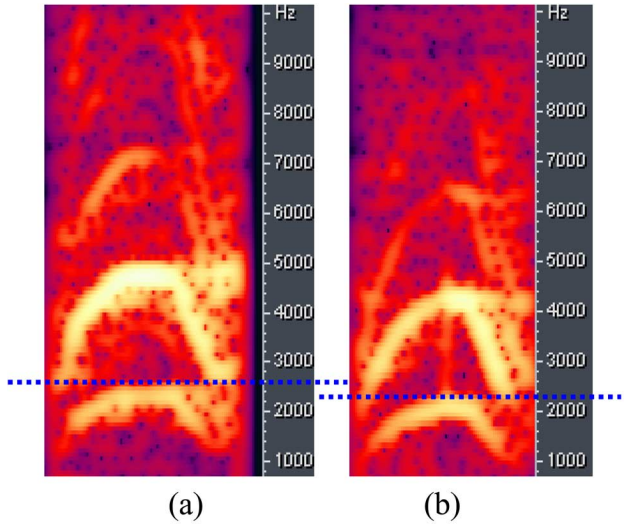


Fig. 3. Spectrograms of Gray-headed Pygmy Woodpecker. (a) Training syllable. (b) Test syllable.

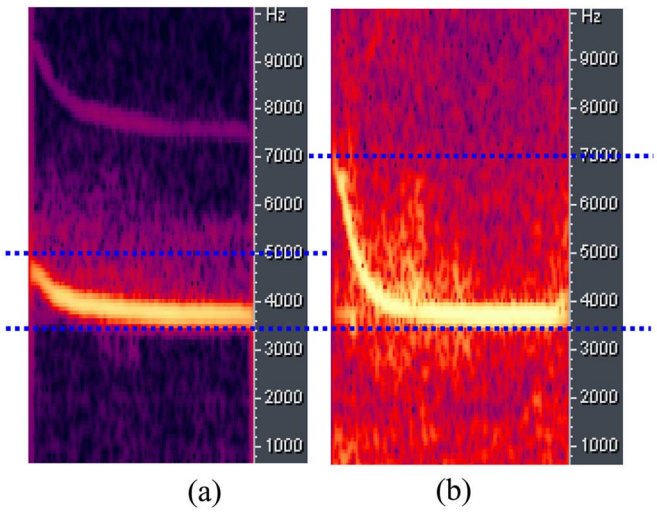


Fig. 4. Spectrograms of Green-backed Tit. (a) Training syllable. (b) Test syllable.

and DTDMFCC, achieves the best classification accuracy among all methods. The best classification accuracy is 84.06% using SDTDMFCC when the PCA threshold $\alpha = 0.97$. Table II summarizes the classification accuracy, the model selected by the SMSS method, and the cluster number selected to model the acoustic variations of each bird species for SDTDMFCC when the PCA threshold $\alpha = 0.97$ is chosen. From this table, we can see that the classification result is acceptable for most bird species.

There are three bird species with their classification accuracy much lower than others, including Gray-headed Pygmy Woodpecker (0%), Green-backed Tit (7.14%), and Gould's Fulvetta (33.33%). Figs. 3–5 show the spectrograms of some training syllables and test syllables of these three bird species. By comparing the spectrograms of the training syllable and test syllable shown in Fig. 3, we can see that the fundamental frequencies of these two syllables are distinct. The frequency of the training syllable is higher than that of the test syllable by about 300 Hz (2500 Hz versus 2200 Hz). In Fig. 4, the test syllable exhibits larger frequency range (3400–7000 Hz) than the training syllable (3400–5000 Hz). From Fig. 5, we can see that the spectrogram of the training syllables exhibits some harmonics at high frequency part, although not very clear. However, the higher frequency part in the spectrogram of the test syllable contains only white noise. Since MFCC-based features are calculated by first computing the energy for each Mel-frequency subband. If the frequency distributions of two syllables are distinct, the calculated MFCC-based features will differ. Thus, the classification results will degrade for such cases as shown in Figs. 3–5.

For each input syllable, the average execution time of each classification module (including feature extraction, PCA, GMM/EVQ, LDA, and classification) during the training phase and the classification phase is shown in Table III. The average duration of a syllable is 0.26 and 0.28 s for the training set and test set, respectively. It can be seen the feature extraction module and the GMM/EVQ clustering module take most of the execution time during the training phase, whereas the feature extraction module takes most of the execution time during the classification phase. The average execution time required to classify each syllable is 0.0971 s, which meets the real-time applications.

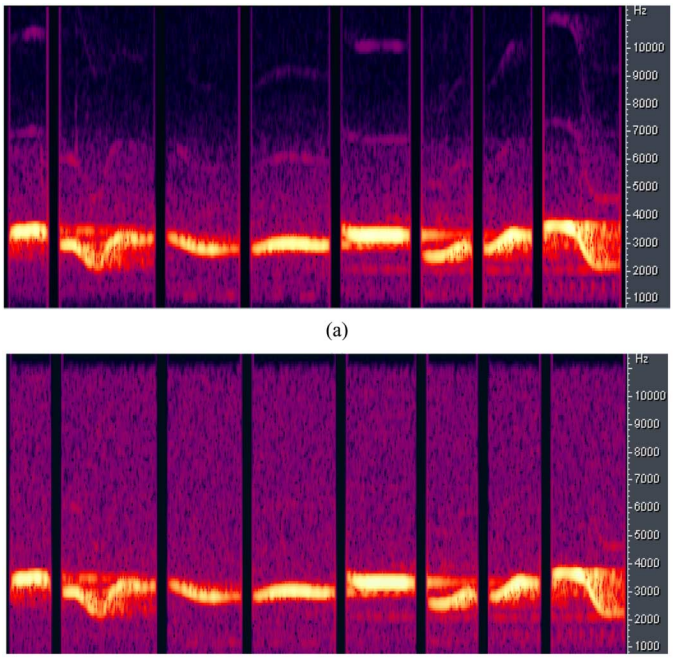


Fig. 5. Spectrograms of Gould's Fulvetta. (a) Training syllables. (b) Test syllables.

TABLE III
AVERAGE EXECUTION TIME REQUIRED FOR THE CLASSIFICATION OF EACH SYLLABLE AS WELL AS EACH CLASSIFICATION MODULE DURING THE TRAINING PHASE AND THE CLASSIFICATION PHASE

	Training Phase	Classification Phase
TDMFCC feature extraction	0.1042	0.0965
PCA	0.0008	0.0002
GMM/EVQ	0.0960	–
LDA	0.0006	0.0003
Classification	–	0.0001
Total	0.2016	0.0971

IV. CONCLUSION

Automatic classification of bird sounds offers a new tool for the classification or differentiation of bird species by their sounds. It is by no means an easy task though it is interesting. This is due to the fact that bird sounds vary among species and even identical species might make many different types of sounds. This paper developed a method for classifying bird species from the sounds they generate. A feature set, TDMFCC and DTDMFCC, are employed as the vocalization features for the classification of each individual syllable segmented from continuous birdsong recordings. In the experiment, the test syllables and training syllables are segmented from different recordings. When both TDMFCC and DTDMFCC are combined together, a classification accuracy of 84.06% can be obtained for the classification of 28 bird species. Based on the results of this paper, it seems that certain bird species are easily to be recognized using the proposed method, whereas some species result in higher classification errors.

Due to the lack of a standard test data set for birdsong recognition, the sets of training and test data are limited. For example, the respective recognition rates for Crested Serpent Eagle, Beavan's Bullfinch, Verreaux's Bush Warbler, and Steere's Liocichla are 100%, 100%, 100%, and 80% based on four, eight, eight, and five test syllables, respectively. These results cannot be considered representative with such a limited data. To make the results more convincing, we will try to collect more training and test recordings in the near future. In addition, a more efficient and robust feature extraction or classification approaches are needed to improve the classification performance under adverse environmental or recording conditions.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments that improved the representation and quality of this paper.

REFERENCES

[1] E. A. Brenowitz, D. Margoliash, and K. M. Nordeen, "An introduction to birdsong and the avian song system," *J. Neurobiol.*, vol. 33, no. 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," *J. Acoust. Soc. Amer.*, 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," *J. Acoust. Soc. Amer.*, 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 *Proc. IEEE Conf. Commun., Power, Comput.*, 1995, vol. 2, pp. 409–414.
[5] A. L. McIlraith and H. C. Card, "A comparison of backpropagation and statistical classifiers for bird identification," in *Proc. IEEE Int. Conf. Neural Netw.*, 1997, vol. 1, pp. 100–104.
[6] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Trans. Signal Process.*, 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 *Proc. Can. Conf. Elect. Comput. Eng.*, 1997, vol. 1, pp. 63–66.
[8] A. Härmä, "Automatic identification of bird species based on sinusoidal modeling of syllables," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2003, vol. 5, pp. 545–548.
[9] A. Härmä and P. Somervuo, "Classification of the harmonic structure in bird vocalization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 5, pp. 701–704.
[10] P. Somervuo and A. Härmä, "Bird song recognition based on syllable pair histograms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 5, pp. 825–828.
[11] P. Somervuo, A. Härmä, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2252–2263, Nov. 2006.
[12] N. C. Singh and F. E. Theunissen, "Modulation spectra of natural sounds and ethological theories of auditory processing," *J. Acoust. Soc. Amer.*, vol. 114, no. 6, pp. 3394–3411, Dec. 2003.
[13] H. Wang, S. Saar, O. Tchernichovski, and P. P. Mitra, "Characterization of birdsong using spectral correlations," presented at the 37th Annu. Meeting Soc. Neurosci., 2007, unpublished.
[14] Y. Ariki, S. Mizuta, M. Magata, and T. Sakai, "Spoken-word recognition using dynamic features analysed by two-dimensional cepstrum," *Proc. Inst. Elect. Eng.*, vol. 136, no. 2, pt. I, pp. 133–140, Apr. 1998.
[15] H. F. Pai and H. C. Wang, "A study of the two-dimensional cepstrum approach for speech recognition," *Comput. Speech Lang.*, vol. 6, no. 4, pp. 361–375, Oct. 1992.
[16] C. T. Lin, H. W. Nein, and J. Y. Hwu, "GA-based noisy speech recognition using two-dimensional cepstrum," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 664–675, Nov. 2000.
[17] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 1, pp. 52–59, Feb. 1986.
[18] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: Wiley, 2000.

[19] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ: Prentice-Hall, 2001.
[20] T. Matsui and S. Furui, "Comparison of text independent speaker recognition methods using VQ distortion and discrete/continuous HMMs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1992, vol. 2, pp. 157–160.
[21] M. Nishida and T. Kawahara, "Speaker model selection based on the Bayesian information criterion applied to unsupervised speaker indexing," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 583–592, Jul. 2005.
[22] S. S. Cheng, H. M. Wang, and H. C. Fu, "A model-selection-based self-splitting Gaussian mixture learning with application to speaker identification," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 17, pp. 2626–2639, 2004.
[23] G. Kolano and P. Regel-Brietzmann, "Combination of vector quantization and Gaussian mixture models for speaker verification with sparse training data," in *Proc. Eur. Conf. Speech Tech. (Eurospeech)*, 1999, pp. 1203–1206.
[24] "CD sound of the mountain IV: The songs of wild birds". Yushan National Park. Yushan, Taiwan, 1995.
[25] "CD sound of the mountain V: The songs of wild birds". Yushan National Park. Yushan, Taiwan, 1996.
[26] C. H. Lee, C. H. Chou, C.-C. Han, and R. Z. Huang, "Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis," *Pattern Recognition Lett.*, vol. 27, no. 2, pp. 93–101, Jan. 2006.

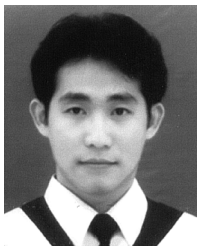
Chang-Hsing Lee was born in Tainan, Taiwan, on July 24, 1968. He received the B.S. and Ph.D. degrees in computer and information science from National Chiao Tung University, Hsinchu, Taiwan, in 1991 and 1995, respectively.

He is currently an Associate Professor in the Department of Computer Science and Information Engineering, Chung Hua University, Hsinchu. His main research interests include audio/sound classification, multimedia information retrieval, and multimedia

Chin-Chuan Han (M'05) received the B.S. degree in computer engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 1989, and the M.S. and a Ph.D. degrees in computer science and electronic engineering from National Central University, Jhongli City, Taiwan, in 1991 and 1994, respectively.

From 1995 to 1998, he was a Postdoctoral Fellow in the Institute of Information Science, Academia Sinica, Taipei, Taiwan. He was an Assistant Research Fellow in the Telecommunication Laboratories, Chunghwa Telecom Co. in 1999. From 2000 to 2004, he worked with the Department of Computer Science and Information Engineering, Chung Hua University, Hsinchu. In 2004, he joined the Department of Computer Science and Information Engineering, National United University, Maio-Li, Taiwan, where he became a Professor in 2007.

Prof. Han is a member of SPIE and IPPR in Taiwan. His research interests are in the areas of face recognition, biometrics authentication, video surveillance, image analysis, computer vision, and pattern recognition.



Ching-Chien Chuang was born in Hsinchu, Taiwan, on April 4, 1978. He received the B.S. and M.S. degrees in computer science and information engineering from Chung Hua University, Hsinchu, in 2005 and 2008, respectively.

His main research interests include image processing and audio processing.