Pattern Recognition 43 (2010) 3846–3852

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines

Jinkui Cheng[a,b], Yuehua Sun[a], Liqiang Ji[a,*]

[a] Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, 1 Beichen West Road, Beijing 100101, China
[b] Graduate University of Chinese Academy of Sciences, 19 Yuquan Road, Beijing 100049, China

ARTICLE INFO

Article history:
Received 5 January 2010
Received in revised form
22 March 2010
Accepted 12 April 2010

Keywords:
Call-independent identification
Gaussian mixture models
Individual recognition
Mel-frequency cepstral coefficients
Passerine

ABSTRACT

Research into acoustic recognition systems for animals has focused on call-dependent and species identification rather than call-independent and individual identification. Here we present a system for automatic call-independent individual recognition using mel-frequency cepstral coefficients and Gaussian mixture models across four passerine species. To our knowledge this is the first application of these techniques to the individual recognition of birds, and the results are promising. Accuracies of 89.1–92.5% were achieved and the acoustic feature and classifier method developed here have excellent potential for individual animal recognition and can be easily applied to other species.

1. Introduction

Many animals use sound to communicate with conspecifics and thus animal vocalizations have evolved to be species specific. Across many taxa, animal calls show individual variation. For example, in fish [1], amphibians [2,3], birds [4,5], and mammals [6,7] animal vocalizations may be individual specific. Given this, species and even individual recognition based on animal vocalizations is possible for many animals and consequently can be utilized as a useful tool in the study and monitoring of animal species.

Automatic species and individual recognition based on acoustic animal call parameters is a challenge. Interest in this field is on the rise and several automatic approaches were recently proposed. One approach gaining results borrows methods from human speech and speaker recognition [8]. First, acoustic features from animal calls recorded in the field are extracted and each call is transformed into a feature vector or set of feature vectors representing salient characteristics. Second, a classifier is trained to distinguish between feature sets. Third, following testing the classifier can be used to classify new recordings as belonging to one of the target classes or to an unknown class [9].

To obtain robust recognition results, effective acoustic features that show greater variation between rather than within species or individuals are needed [10]. These acoustic features can be classified into two classes: statistical and non-statistical. Statistical features include mean fundamental frequency, maximum fundamental frequency, minimum fundamental frequency, fundamental range, syllable energy, syllable duration, zero-crossing rate and signal bandwidth [11,12]. Long-term averages of these statistical features have been utilized in machine-learning algorithms that have successfully identified different bird and frog species [13,14]. Statistical call features can also be used to identify individuals, although long-term averages discard a great deal of individual information and condense call characteristics [15]. Weary et al. [16] achieved call-dependent recognition accuracies of between 69% and 80% in grey tits (*Parus afer*); and Amazonian manatees (*Trichechus inunguis*) can be differentiated based on individual differences in fundamental frequency and signal duration [11].

Non-statistical features such as linear prediction coefficients (LPCs) [17] and mel-frequency cepstral coefficients (MFCCs) [18,19] are common in human speech and speaker recognition systems. Applying these features to species identification have yielded results across a variety of taxa including frogs, crickets [20] and birds [21,22]. The application of non-statistical features to individual recognition has proven to be more difficult and results are varied. In African elephants (*Loxodonta africana*), 83% individual recognition accuracy was achieved [23] and in Norwegian ortolan bunting (*Emberiza hortulana*) 80–95% of individuals were identified correctly [24]. In general, models based on non-statistical features are of greater accuracy, stability and repeatability.

[*] Corresponding author. Tel.: +86 10 64807129; fax: +86 10 64807099.
E-mail address: ji@ioz.ac.cn (L. Ji).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.04.026

Feature classification methods developed for human speech recognition have been applied to species and individual recognition in animals. These methods include dynamic time warping (DTW) [25], sinusoidal modeling of syllables [26], self-organizing maps [27,28], linear discriminant analysis (LDA) [20], artificial neural network (ANN) [10,21], sport vector machine (SVM) [13,14], Gaussian mixture models (GMM) [9] and hidden markov models (HMM) [22,24]. In speech and speaker recognition, the type of classifier selected depends on the task required [29] so chosen classifiers for species and individual recognition in animals must be carefully considered.

The majority of research into animal recognition is call dependent and focused predominantly on species identification rather than individual identification. Call-dependent systems are limited because they rely on recognition techniques that can compare only a single call type within and between individuals and thus significantly limit the range of species and situations in which they can be applied. Achieving call-independent recognition is more challenging, but enables recognition regardless of the call type produced [30]. Here, we aim to construct an automatic call-independent recognition system and test the ability of GMM to achieve this for four passerines: Gansu leaf warbler (*Phylloscopus kansuensis*), Chinese leaf warbler (*Phylloscopus yunnanensis*), Hume's warbler (*Phylloscopus humei*) and Chinese bulbul (*Pycnonotus sinensis*).

2. Method

The architecture of our acoustic-driven individual recognition system for birds can be divided into three modules: signal preprocessing, feature extraction, and classification (see Fig. 1).

2.1. Data set

One song type was recorded from Hume's warbler ($N=10$ birds) and two song types were recorded from Gansu leaf warblers ($N=5$), Chinese leaf warblers ($N=9$) and Chinese bulbuls ($N=10$) were recorded. There are strong distinctions between the songs of these species (see Fig. 2). Chinese leaf warblers were recorded from Taibaishan National Nature Reserve (33Chinese–343Chinese 107Chinese–107Chinese le Gansu leaf warblers and Hume's warblers were recorded from Lianhuashan National Nature Reserve (34nsu –344nsu l 103nsu –103nsu leaf warblers and Hume's warblers were recorded from Lianhuashan National Nature Reserve or call-independent training and testing for a) Hume's warbler, b) Chinese leaf WarWM-D6c professional recorder (Sony Corporation, Tokyo, Japan) with a directional microphone (Sennheiser, Wedemark, Germany) placed 2–8 m from a singing bird. Recordings were converted to a digital medium at 22.05 kHz sampling frequency and saved in 8-bit wave format using Batsound v3.10 (Pettersson Elektronik AB, Uppsala, Sweden).

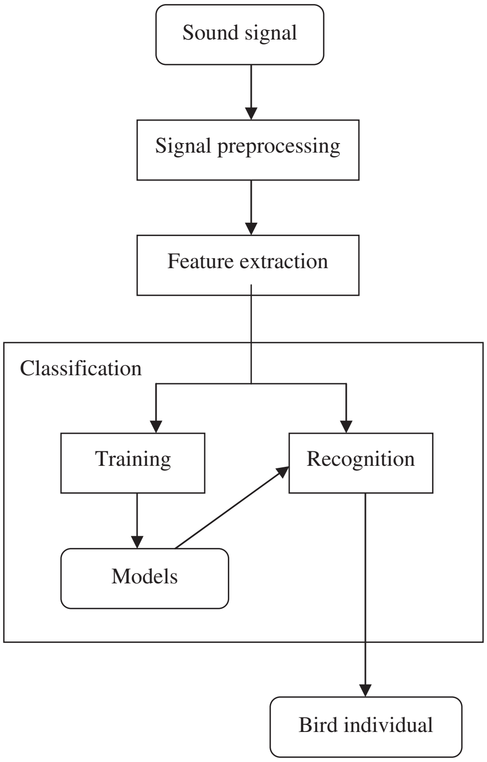

Fig. 1. Architecture of our individual recognition system.

2.2. Feature extraction

2.2.1. Sound signal preprocessing

Bird song is typically divided into four hierarchical levels of notes, syllables, phrases, and song [31]. Of these, syllables are the most elementary building blocks and suitable for species and individual recognition as variation in this aspect of song is neither excessive not leads to model instability [26,32]. Prior to feature extraction syllables must be segmented; here we used an iterative time-domain algorithm [33] following the protocols of Huang et al. [14]. Once segmented, sound signals (now consisting of syllables) were divided into two sets to train the classifier and test the classifier (Table 1). Humans generate speech by exciting the vocal cords and the high frequencies of human speech are weakened during the production. Therefore, there is a need to enhance the high frequencies by a digital filter during pre-emphasized processing. Bird sounds are generated mainly by the syrinx but sound generation in birds is similar to that in humans [21]. Bird sound signals were pre-emphasized before extracting features by a digital filter described by the formula

$$H(Z)=1-\mu z^{-1} \tag{1}$$

where μ is 0.95.

The signal was then divided into a set of overlapping frames with a frame size of 400 samples, and overlapping size of 200 samples for each pair of successive frames. To reduce discontinuity on both ends of a frame each frame was multiplied by the Hamming window

$$S[n]=s[n]w[n], \quad 0 \le n \le N-1 \tag{2}$$

where $S[n]$ is the output signal, $s[n]$ is the signal denoting the input syllable, $w[n]$ is the Hamming window function and N is 512.

$$w[n]=0.54-0.46\cos(2\pi n/N-1), \quad 0 \le n \le N-1 \tag{3}$$

We then took the discrete Fourier transform of each frame using the Fast Fourier Transform (FFT).

$$X[k]=\sum_{n=0}^{N-1} s[n]\exp\{-2jk\pi n/N\}, \quad 0 \le k \le N-1 \tag{4}$$

where $X[k]$ is the output signal and $s[n]$ is the input signal denoting the signal obtained above.

2.2.2. MFCCs extraction

After signal preprocessing, the MFCCs features can be extracted from each frame. In studies of speech recognition, the MFCCs and LPCs are commonly used; however the MFCCs perform better than others in recognition accuracy [34–36] and have been widely used for bird song recognition [25].

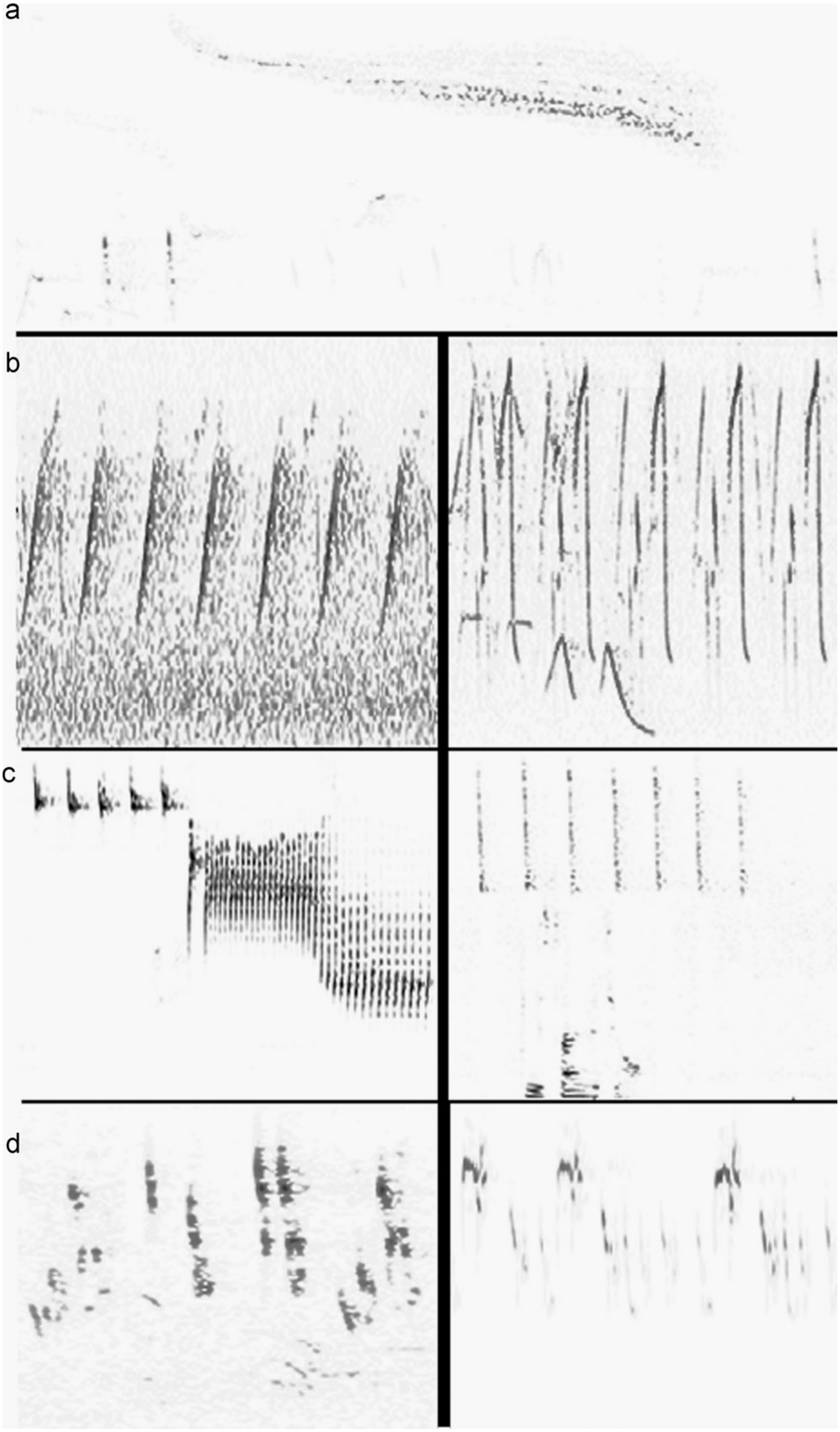

Fig. 2. Example of the spectrograms of different call types used for call-independent training and testing for a) Hume's warbler, b) Chinese leaf Warbler, c) Gansu leaf warbler and d) Chinese bulbul.

Human auditory perception does not follow a linear scale and the perception of some frequencies are greatly influenced by energy in the critical band of frequencies around them. The bandwidth of the critical band varies with the perceived frequency. The advantage of this is that the system is capable of being immune to noise and easily warps frequencies to a

Table 1
The segmenting results of each individual sounds

Species	Data set	Syllable type numbers	Syllable numbers of each individual									
			1	2	3	4	5	6	7	8	9	10
Hume's warbler	Training set	1	7	8	10	6	10	7	9	7	5	7
	Testing set	1	4	4	5	4	4	4	6	4	3	4
Chinese leaf warbler	Training set	3	72	56	120	112	104	48	52	88	76	
	Testing set	3	34	12	52	80	42	28	22	32	40	
Gansu leaf warbler	Training set	3	35	60	40	40	30					
	Testing set	3	20	40	14	18	16					
Chinese bulbul	Training set	6	24	20	62	40	45	67	50	72	53	
	Testing set	6	13	15	30	27	30	34	20	32	24	

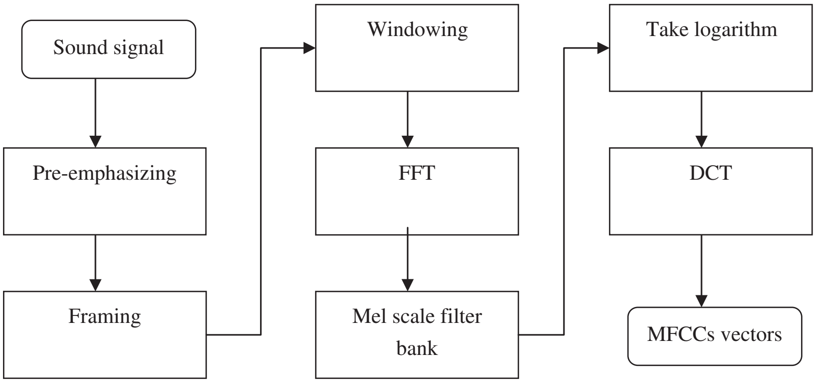

Fig. 3. MFCCs vector extraction.

non-uniform scale, such as the mel scale [37]. Thus, for a signal with an actual frequency, f (measured in Hz), a subjective pitch is measured on the mel scale. The relationship between the real frequency scale and the mel-frequency scale can be modeled as

$$F_{mel} = 2595\log\log(1+F_{rel}/700) \tag{5}$$

where, F_{mel} is the mel frequency and F_{rel} is the real frequency.

The mel scale was used to transform the power spectrum to compute a mel-warped spectrum. In order to simplify the spectrum without significant loss of data, Fourier transformed signals were passed through a set of band-pass filters. The filters were assumed to be triangular and half overlapping, with center frequencies spaced equally apart on the mel scale but non-uniformly distributed on the real scale (see [37,38] for graphical representation). Each filter in the bank was multiplied by the spectrum so that only a single value of magnitude per filter was returned. This value can be calculated by the formula

$$m(l) = \sum_{k=o(l)}^{k=h(l)} w_l(k)|X[k]|, \quad l = 1,2,\ldots,L \tag{6}$$

where $m(l)$ is the output result of each filter, L the number of the triangular filters, in our experiment L is 24. $X[k]$ is the input signal denoting the signal obtained in the sound signal preprocessing step.

This value reflected the sum of amplitudes in a particular filter band and thus reduced precision to the level of the human auditory system. The modified spectrum consisted of the output power of these filters. A logarithm of the mel spectrum coefficients was then taken to compress the coefficients above 1000 Hz and also to compress the magnitude with low frequencies. In our final step, the MFCC coefficients were obtained by taking a discrete cosine transform (DCT) following [39]

$$C_{mfcc}(i) = \sqrt{2/N}\sum_{l=1}^{L}\log m(l)\cos((l-1/2)i\pi/L), \quad 1 \le i \le L \tag{7}$$

where $C_{mfcc}(i)$ is the value of the i-th dimension of the MFCCs feature vector extracted from the frame. Here, the MFCCs feature vectors are in 23 dimensions (see Fig. 3)

2.3. Classification

Research into speech recognition has shown that probabilistic models provide a better model of acoustic speech events and framework for dealing with noise and channel degradation than non-probabilistic models [40]. We selected GMM as a representation of bird identity because the Gaussian components of the GMM can represent general bird-dependent spectral shapes and because of its ability to model arbitrary densities [41,42]. Multivariate Gaussians have a well-recorded history of serving as good radial-bases functions for functional approximation [43], functional approximation using the Gaussian parametric probability density functions inherently solves the problem of over-fitting due its smooth and multivariate nature [44]. A Gaussian mixture density is a weighted sum of M component densities, as given by the equation

$$p(x|\lambda) = \sum_{i=1}^{M} p_i b_i(x), \quad i = 1,2,\ldots,M \tag{8}$$

where x is a D-dimensional random vector, $b_i(x)$ is the component densities and p_i is the mix weights. Each component density is a D-variate Gaussian function defined as

$$b_i(x) = 1/\left((2\pi)^{D/2}\left|\sum i\right|^{1/2}\right)\exp\left\{-1/2(x-\mu_i)\sum i^{-1}(x-\mu_i)\right\} \tag{9}$$

where, μ_i is the mean vector and $\sum i$ is the covariance matrix. The mixture weights satisfy the constraint that

$$\sum_{i=1}^{M} p_i = 1 \tag{10}$$

The complete GMM can be parameterized by the mean vectors, covariance matrices and mixture weighs from all component densities. The parameters of the GMM, λ, are denoted as

$$\lambda = \left\{ p_i, \mu_i, \sum i \right\}, \quad i = 1,2,\ldots,M \tag{11}$$

where, M is the number of the components of the GMM.

For bird identification, each bird is represented by a GMM. In order to estimate the parameters of the GMM, λ, which best matches the distribution of the training feature vectors, we used maximum likelihood (ML) estimation [42]. So for a set of T training vectors $X = \{x_1, x_2, \ldots, x_t\}$, the GMM likelihood can be written as

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \tag{12}$$

The parameters of the GMM were obtained by maximizing the likelihood function and obtained iteratively using the expectation–maximization (EM) algorithm [45,46]. On each EM iteration, the following re-estimation formulas were used and guaranteed a monotonic increase in model's likelihood value.

Mixture Weights

$$p_i = 1/T \sum_{t=1}^{T} p(i|x_t,\lambda) \tag{13}$$

Means

$$\mu_i = \sum_{t=1}^{T} p(i|x_t,\lambda) x_t \Big/ \sum_{t=1}^{T} p(i|x_t,\lambda) \tag{14}$$

Covariances

$$\sum i = \sum_{=1} p(i|x_t,\lambda)(x_t - \mu_t)(x_t - \mu_i)^T \Big/ \sum_{t=1}^{T} p(i|x_t,\lambda) \tag{15}$$

A posteriori probability for acoustic class is given by the following equation:

$$p(i|x_t,\lambda) = p_i b_i(x_t) \Big/ \sum_{k=1}^{M} p_k b_k(x_t) \tag{16}$$

We calculated the mean value for each dimension of the feature vector and complete covariance matrix using all vectors from each individual. We then initiated the initial mean vector and the initial covariance matrix for each individual using the mean values and the covariance matrix. The initial mixture weights for each component of the GMM were set as equal. Then the GMMs for each individual were trained using the EM algorithm. For bird identification, a group of k individual birds is represented by a set of GMMs: $\lambda_1, \lambda_2, \ldots, \lambda_k$. The objective is to find a model, M, which has the maximum a posteriori probability (MAP) for a given observation sequence $X = \{x_1, x_2, \ldots, x_t\}$

Formally,

$$M = \mathrm{argmax}\Pr(\lambda_k/X) = \arg\max p(x|\lambda_k)\Pr(\lambda_k)/p(X), \quad 1 \le k \le K \tag{17}$$

where, K is the number of bird individuals, the second equation is due to Bayes' rule. Assuming equally likely the models (i.e., $\Pr(\lambda_k) = 1/K$) and noticing that $p(X)$ is the same for all bird models. So the classification rule can be simplified to

$$M = \arg\max p(X|\lambda_k) \tag{18}$$

Using logarithms and the independence between observations, the individual recognition system computes

$$M = \arg\max \sum_{t=1}^{T} \log p(x_t|\lambda_k)$$

in which, QUOTE $p(x_t|\lambda_k)$ is given in formula (8) [42].

3. Results

In order to investigate individual recognition performance of the GMM with respect to the number of component densities per model, each bird was modeled using 4, 8, 16, 32, 48 and 64 component GMM with a complete covariance matrix. To determine the length of the testing signal, each model was tested using 23-dimensional mel-frequency cepstral vectors corresponding to a 1 and 2 s testing signal. Identification performances for different model orders and testing signal lengths are shown in Table 2. From our results is it clear that the GMMs are sensitive to the number of mixture components for both 1 and 2 s testing signals, and GMMs with the best results across different species had different mixture components. However, model order is not directly proportional to the identification accuracy of the model. Optimal model order when using a 1 s testing signal was found to be 16 (for Gansu leaf warbler), 16 (Chinese leaf warbler), 32 (Hume's warbler) and 4 (Chinese bulbul) and for a 2 s signal 16 (Gansu leaf warbler) , 16 (Chinese leaf warbler), 8 (Hume's warbler) and 4 (Chinese bulbul). Table 2 indicates that when model order exceeds optimal model order recognition results decrease. We posit that because bird sounds are not complex, an appropriate model order and not the largest model order is required. When model order exceeds optimal model order the result is over-fitting. Our results also show that identification results using a 2 s testing signal were more accurate than using a testing signal of 1 s. Our final recognition system therefore adopted a signal testing length of 2 s and associated mixture components of the GMM.

Call-independent recognition in the four passerine species using the MFCCs and GMM achieved good recognition accuracy. The accuracy was 90% for Gansu leaf warbler, 89.1% for Chinese leaf warbler, 92.5% for Hume's warbler, and 90.2% for Chinese bulbul. Confusion matrices based on identification results are shown in Tables 3–6 where the first column describes the number of the testing vectors per individual.

Table 2
GMM identification performance across different model orders and test lengths.

Test length	Model order	Species			
		Hume's warbler (%)	Chinese leaf warbler (%)	Gansu leaf warbler (%)	Chinese bulbul (%)
1 s	M=4	75.9	76.1	85.3	78.5
	M=8	78.4	79.1	87.3	74.8
	M=16	79.3	82.1	91.2	72.0
	M=32	81.1	79.1	76.5	69.2
	M=48	66.7	81.3	89.2	52.3
	M=64	64.0	66.4	77.5	29.9
2 s	M=4	88.9	81.3	83.7	90.2
	M=8	92.5	85.5	87.8	82.4
	M=16	88.7	89.1	90.0	80.4
	M=32	83.0	82.8	77.6	72.6
	M=48	75.5	82.8	87.8	52.9
	M=64	67.9	67.2	75.5	32.7

Table 3
Confusion matrix of the recognition results in the Gansu leaf warbler.

Number		1	2	3	4	5
10	1	6	4	0	0	0
18	2	0	18	0	0	0
7	3	0	0	7	0	0
8	4	1	0	0	7	0
7	5	0	0	0	0	7

Table 4
Confusion matrix of the recognition results in Chinese leaf warbler.

Number		1	2	3	4	5	6	7	8	9
7	1	4	0	0	3	0	0	0	0	0
2	2	0	2	0	0	0	0	0	0	0
4	3	0	0	2	2	0	0	0	0	0
16	4	0	0	0	16	0	0	0	0	0
9	5	0	0	0	0	9	0	0	0	0
6	6	0	0	0	0	0	6	0	0	0
4	7	0	0	0	0	0	0	4	0	0
7	8	0	0	0	0	0	0	2	5	0
9	9	0	0	0	0	0	0	0	0	9

Table 5
Confusion matrix of the recognition results in Hume's warbler.

Number		1	2	3	4	5	6	7	8	9	10
4	1	4	0	0	0	0	0	0	0	0	
5	2	0	5	0	0	0	0	0	0	0	
6	3	0	0	5	0	0	0	0	0	0	1
4	4	0	0	0	4	0	0	0	0	0	
6	5	0	0	0	0	6	0	0	0	0	
5	6	0	0	0	0	0	5	0	0	0	
7	7	0	0	0	0	0	0	6	0	1	
6	8	0	0	0	0	0	0	0	6	0	
4	9	1	0	0	0	0	0	0	0	3	
6	10	1	0	0	0	0	0	0	0	0	5

Table 6
Confusion matrix of the recognition results in Chinese bulbul.

Number		1	2	3	4	5	6	7	8	9	10
2	1	2	0	0	0	0	0	0	0	0	0
2	2	0	2	0	0	0	0	0	0	0	0
6	3	0	0	6	0	0	0	0	0	0	0
4	4	0	0	0	3	1	0	0	0	0	0
6	5	0	0	0	0	6	0	0	0	0	0
6	6	0	0	0	0	0	6	0	0	0	0
5	7	0	0	0	0	0	0	5	0	0	0
6	8	0	0	0	0	0	0	0	3	3	0
7	9	0	0	0	0	0	0	0	0	7	0
7	10	0	0	0	0	0	0	0	0	1	6

4. Discussion

Our avian call-independent recognition system uses methods of feature extraction and classification adapted from human speech and speaker recognition technology. This system optimizes the number of mixture components of the GMM for each passerine species. However, the model will continue to be improved by modifying the methods to better suit bird song or by incorporating individual-specific information and optimizing the GMM for each individual [47].

Call recordings used here were made in the field and noise may therefore be a potential adverse factor that prevented us from achieving higher identification accuracies. For example, the recordings used when testing a classifier had noise profiles different from those used in classifier training. Noise is also a major challenge in human speech and speaker recognition applications [48] and can arise from a variety of sources such an ambient noise, reverberations, channel interference and microphone distortions [10]. The recordings used here did contain background noise but the noise was relatively small as data was collected prior to sunrise and with a directional microphone six meters from the focal bird. During model development we excluded recordings that contained large levels of background noise. Although the recordings we used for training and testing contained different background noise we obtained good results demonstrating that the system constructed here has some resilience to noise. Developing strong systems immune to different sources of noise is a challenge facing future research in this field.

Here we present an acoustic system for automatic call-independent individual recognition using the MFCCs and GMM in birds. To our knowledge this is the first application of the GMM techniques to bird individual recognition, and the results are promising. Accuracies of 89.1%–92.5% were achieved across four passerine families. This indicates that the acoustic feature (MFCCs) and classifier (GMM) method developed have excellent potential for individual animal recognition and can be easily applied to other species. Call-independent identification remains a challenging area of research, but can be applied to all species regardless of the amount of song sharing or temporal change in vocal repertoires. Recognition accuracy may be improved by optimizing the GMM for each individual or by incorporating individual-specific information and further research will most likely adopt this strategy.

Acknowledgements

The authors thank staff at the Avian Ecology Group, Key Laboratory of Animal Ecology and Conservation Biology, Ornithological Research Group, Key Laboratory of Zoological Systematics and Evolution (all of the Institute of Zoology) for recordings of birds. Thanks to Jinlin Li, Nan Lu and Xiaoying Xing. We are also grateful to members of the Biodiversity Informatics Group for their support, especially Jiangning Wang and Bengui Xie.

References

[1] J.D. Crawford, A.P. Cook, A.S. Heberlein, Bioacoustic behavior of African fishes (Mormyridae): potential cues for species and individual recognition in Pollimyrus, Journal of the Acoustical Society of America 102 (1997) 1200–1212.

[2] M.A. Bee, C.E. Kozich, K.J. Blackwell, H.C. Gerhardt, Individual variation in advertisement calls of territorial male green frogs, *Rana clamitans*: implications for individual discrimination, Ethology 107 (2001) 65–84.

[3] D. Rogers, Intraspecific Variation in the Acoustic Signals of Birds and Frogs: Implication for the Acoustic Identification of Individuals, University of Adelaide, South Australia, 2002.

[4] A.M.R. Terry, T.M. Peake, P.K. McGregor, The role of vocal individuality in conservation, Frontiers in Zoology 2 (2005) 10.

[5] D.T. Blumstein, O. Munos, Individual, age and sex-specific information is contained in yellow-bellied marmot alarm calls, Animal Behaviour 69 (2005) 353–361.

[6] S.K. Darden, T. Dabelsteen, S.B. Pedersen, A potential tool for swift fox (*Vulpes velox*) conservation: individuality of long-range barking sequences, Journal of Mammalogy 84 (2003) 1417–1427.

[7] K.H. Frommolt, M.E. Goltsman, D.W. MacDonald, Barking foxes, *Alopex lagopus*: field experiments in individual recognition in a territorial mammal, Animal Behaviour 65 (2003) 509–518.

[8] E.J.S. Fox, J.D. Roberts, M. Bennamoun, Text-independent speaker identification in birds, in: Proceedings of the Interspeech 2006 and Ninth International Conference on Spoken Language Processing, vols. 1–5, 2006, pp. 2122–2125.

[9] M.A. Roch, M.S. Soldevilla, J.C. Burtenshaw, E.E. Henderson, J.A. Hildebrand, Gaussian mixture model classification of odontocetes in the Southern

California Bight and the Gulf of California, Journal of the Acoustical Society of America 121 (2007) 1737–1748.

[10] E.J.S. Fox, A new perspective on acoustic individual recognition in animals with limited call sharing or changing repertoires, Animal Behaviour 75 (2008) 1187–1194.

[11] R.S. Sousa-Lima, A.P. Paglia, G.A.B. Da Fonseca, Signature information and individual recognition in the isolation calls of *Amazonian manatees*, *Trichechus inunguis* (Mammalia : Sirenia), Animal Behaviour 63 (2002) 301–310.

[12] C. Molnar, F. Kaplan, P. Roy, F. Pachet, P. Pongracz, A. Doka, A. Miklosi, Classification of dog barks: a machine learning approach, Animal Cognition 11 (2008) 389–400.

[13] M.A. Acevedo, C.J. Corrada-Bravo, H. Corrada-Bravo, L.J. Villanueva-Rivera, T.M. Aide, Automated classification of bird and amphibian calls using machine learning: a comparison of methods, Ecological Informatics 4 (2009) 206–214.

[14] C.J. Huang, Y.J. Yang, D.X. Yang, Y.J. Chen, Frog classification using machine learning techniques, Expert Systems with Applications 36 (2009) 3737–3743.

[15] D.A. Reynolds, Large population speaker identification using clean and telephone speech, IEEE Signal Processing Letters 2 (1995) 46–48.

[16] D.M. Weary, K.J. Norris, J.B. Falls, Song features birds use to identify individuals, Auk 107 (1990) 623–625.

[17] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, NJ, 1993.

[18] J.W. Picone, Signal modeling techniques in speech recognition, Proceedings of the IEEE 81 (1993) 1215–1247.

[19] J.P. Campbell, Speaker recognition: a tutorial, Proceedings of the IEEE 85 (1997) 1437–1462.

[20] C.H. Lee, C.H. Chou, C.C. Han, R.Z. Huang, Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis, Pattern Recognition Letters 27 (2006) 93–101.

[21] C.F. Juang, T.M. Chen, Birdsong recognition using prediction-based recurrent neural fuzzy networks, Neurocomputing 71 (2007) 121–130.

[22] V.M. Trifa, A.N.G. Kirschel, C.E. Taylor, E.E. Vallejo, Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models, Journal of the Acoustical Society of America 123 (2008) 2424–2431.

[23] P.J. Clemins, M.T. Johnson, K.M. Leong, A. Savage, Automatic classification and speaker identification of African elephant (*Loxodonta africana*) vocalizations, Journal of the Acoustical Society of America 117 (2005) 956–963.

[24] M.B. Trawicki, M.T. Johnson, T.S. Osiejuk, Automatic song-type classification and speaker identification of Norwegian Ortolan Bunting (Emberiza hortulana) vocalizations, in: Proceeding of the 2005 IEEE Workshop on Machine Learning for Signal Processing (MLSP), (2005) 277-282.

[25] J.A. Kogan, 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 103 (1998) 2185–2196.

[26] A. Harma, Automatic identification of bird species based on sinusoidal modeling of syllables, in: Proceeding of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. V, 2003, pp. 545–548.

[27] P. Somervuo, Competing hidden Markov models on the Self-Organizing Map, Ijcnn 2000, in: Proceedings of the IEEE-Inns-Enns International Joint Conference on Neural Networks, vol. Iii, 2000, pp. 169–174.

[28] E.E. Vallejo, M.L. Cody, C.E. Taylor, Unsupervised acoustic classification of bird species using hierarchical self-organizing maps, Progress in Artificial Life, Proceedings 4828 (2007) 212–221.

[29] R.P. Ramachandran, K.R. Farrell, R. Ramachandran, R.J. Mammone, Speaker recognition—general and data fusion classifier approaches methods, Pattern Recognition 35 (2002) 2801–2821.

[30] E.J.S. Fox, J.D. Roberts, M. Bennamoun, Call-independent individual identification in birds, Bioacoustics—the International Journal of Animal Sound and Its Recording 18 (2008) 51–67.

[31] W.A. Searcy, Bird song: biological themes and variations—Catchpole,CK, Slater,PJB, Animal Behaviour 51 (1996) 492–493.

[32] S.E. Anderson, A.S. Dave, D. Margoliash, Template-based automatic recognition of birdsong syllables from continuous recordings, Journal of the Acoustical Society of America 100 (1996) 1209–1219.

[33] S. Fagerlund, Bird species recognition using support vector machines, Eurasip Journal on Advances in Signal Processing (2007).

[34] R. Vergin, D. OShaughnessy, V. Gupta, Compensated Mel frequency cepstrum coefficients, in: Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, Conference Proceedings, vols. 1–6, 1996, pp. 323–326.

[35] S.B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics Speech and Signal Processing 28 (1980) 357–366.

[36] J.P. Openshaw, Z.P. Sun, J.S. Mason, A Comparison of composite features under degraded speech in speaker recognition, Icassp-93, in: Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, vols. 1–5, 1993, pp. B371–B374.

[37] J.C. Wang, J.F. Wang, Y.S. Weng, Chip design of MFCC extraction for speech recognition, Integration—the Vlsi Journal 32 (2002) 111–131.

[38] S. Panchapagesan, A. Alwan, Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC, Computer Speech and Language 23 (2009) 42–64.

[39] S. Chauhan, P. Wang, C.S. Lim, V. Anantharaman, A computer-aided MFCC-based HMM system for automatic auscultation, Computers in Biology and Medicine 38 (2008) 221–233.

[40] J.S.F. Elizabeth, A new perspective on acoustic individual recognition in animals with limited call sharing or changing repertoires, Animal Behaviour 75 (2008) 1187–1194.

[41] R.A. Redner, H.F. Walker, Mixture densities, maximum-likelihood and the em algorithm, Siam Review 26 (1984) 195–237.

[42] D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using gaussian mixture speaker models, IEEE Transactions on Speech and Audio Processing 3 (1995) 72–83.

[43] H.W. Sorenson, D.L. Alspach, Recursive Bayesian estimation using Gaussian sums, Automatica 7 (1971) 465.

[44] A.D. Subramaniam, B.D. Rao, PDF optimized parametric vector quantization of speech line spectral frequencies, IEEE Transactions on Speech and Audio Processing 11 (2003) 130–142.

[45] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via Em algorithm, Journal of the Royal Statistical Society Series B-Methodological 39 (1977) 1–38.

[46] K.Y. Lee, Local fuzzy PCA based GMM with dimension reduction on speaker identification, Pattern Recognition Letters 25 (2004) 1811–1817.

[47] P.J. Clemins, M.T. Johnson, Generalized perceptual linear prediction features for animal vocalization analysis, Journal of the Acoustical Society of America 120 (2006) 527–534.

[48] B.H. Juang, Speech recognition in adverse environments, Computer Speech and Language 5 (1991) 275–294.

About the Author—JINKUI CHENG received his BSc in Bioinformation from Hebei University, Hebei, China in 2007. He is currently completing PhD at the Graduate University and Institute of Zoology of the Chinese Academy of Sciences. His research interests are animal vocalization processing and recognition.

About the Author—YUEHUA SUN started working in northwest China in 1995, and established a long-term study on the endemic Chinese grouse (*Bonasa sewerzowi*), including population biology, behaviour, landscape ecology and conservation biology. He now focuses on the endemic birds of forests of the Qinghai–Tibet Plateau. Current projects include studies on the Chinese grouse, Sichuan jay and other endemic birds.

About the Author—LIQIANG JI began his research in 1985. His research focus is biodiversity informatics including key techniques and the methodology of collection, processing, publishing and sharing of biodiversity data and information. He is also interested in developing methods and software tools for biodiversity assessments, planning, and implementation of biodiversity information systems and databases.