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).

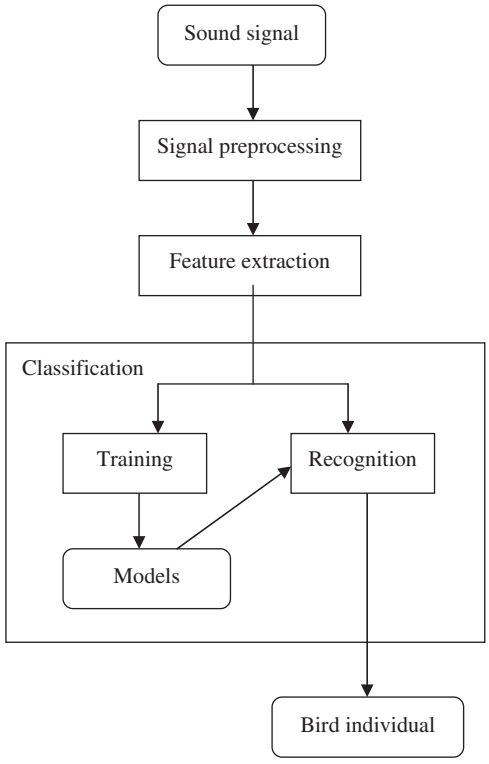

Fig. 1. Architecture of our individual recognition system.

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].