Ecological Informatics 32 (2016) 134–144

Contents lists available at ScienceDirect

Ecological Informatics

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

Adaptive frequency scaled wavelet packet decomposition for frog call classification

Jie Xie *, Michael Towsey, Jinglan Zhang, Paul Roe

Electrical Engineering and Computer Science School, Queensland University of Technology, Brisbane, Australia

ARTICLE INFO

Article history:
Received 8 September 2015
Received in revised form 26 January 2016
Accepted 27 January 2016
Available online 4 February 2016

Keywords:
Frog call classification
Spectral peak track
Adaptive frequency scaled wavelet packet decomposition
k-means clustering
k-nearest neighbour
Support vector machine

ABSTRACT

Environmental changes have put great pressure on biological systems leading to the rapid decline of biodiversity. To monitor this change and protect biodiversity, animal vocalizations have been widely explored by the aid of deploying acoustic sensors in the field. Consequently, large volumes of acoustic data are collected. However, traditional manual methods that require ecologists to physically visit sites to collect biodiversity data are both costly and time consuming. Therefore it is essential to develop new semi-automated and automated methods to identify species in automated audio recordings. In this study, a novel feature extraction method based on wavelet packet decomposition is proposed for frog call classification. After syllable segmentation, the advertisement call of each frog syllable is represented by a spectral peak track, from which track duration, dominant frequency and oscillation rate are calculated. Then, a k-means clustering algorithm is applied to the dominant frequency, and the centroids of clustering results are used to generate the frequency scale for wavelet packet decomposition (WPD). Next, a new feature set named *adaptive frequency scaled wavelet packet decomposition sub-band cepstral coefficients* is extracted by performing WPD on the windowed frog calls. Furthermore, the statistics of all feature vectors over each windowed signal are calculated for producing the final feature set. Finally, two well-known classifiers, a k-nearest neighbour classifier and a support vector machine classifier, are used for classification. In our experiments, we use two different datasets from Queensland, Australia (18 frog species from commercial recordings and field recordings of 8 frog species from James Cook University recordings). The weighted classification accuracy with our proposed method is 99.5% and 97.4% for 18 frog species and 8 frog species respectively, which outperforms all other comparable methods.

1. Introduction

During the past decades, a rapid decline in frog biodiversity has been noted worldwide. There are many reasons for this decline, including habitat destruction (Clauzel et al., 2015), invasive species (Shine, 2014), and climate change (Garcia et al., 2014). Researchers investigate frogs to retain their biodiversity and develop effective protection strategies. Due to the development of acoustic sensor techniques, many sensors have been widely deployed for monitoring biodiversity, which produces large volumes of acoustic data (Wimmer et al., 2013). Compared with the traditional manual methods that require ecologists to physically visit sites for collecting biodiversity data, acoustic sensors can help collect audio data over larger spatio-temporal scales (Wimmer et al., 2010; Gage and Axel, 2014) Since several gigabytes of compressed data can be generated by an acoustic sensor per day, enabling automating species identification in acoustic data sets has become important (Zhang et al., 2013).

In recent years, acoustic data has been studied for the recognition and classification of animal calls by many researchers. Almost all the recognition and classification methods consist of four parts: pre-processing, syllable segmentation, feature extraction, and recognition or classification.

Frog call classification has been addressed in several papers. Huang et al. (2009) extracted spectral centroid, signal bandwidth, and threshold crossing rate from each segmented frog syllable. Then, two classifiers, k-nearest neighbour (k-NN) classifier and support vector machine (SVM), were used for classification. However, signal bandwidth and threshold crossing rate are very sensitive to the background noise, which results in low classification accuracy in noisy environments. Han et al. (2011) introduced spectral centroid, Shannon entropy and Rènyi entropy to classify frog calls with a k-NN classifier. Chen et al. (2012) first calculated syllable length for pre-classification of frog calls based on segmented frog syllables. Then, a multi-stage average spectrum was calculated for automatic recognition based on template matching. However, extracting features based on the Fourier transform has a tradeoff between time and frequency resolution, which restricts the discriminability of the features. Bedoya et al. (2014) proposed an automatic recognition system for frog calls based on the Mel-frequency cepstral coefficients (MFCCs) and a fuzzy classifier. However, MFCCs

* Corresponding author.
E-mail address: xiej8734@gmail.com (J. Xie).

http://dx.doi.org/10.1016/j.ecoinf.2016.01.007
1574-9541/

are designed for the human auditory system, and might be not suitable for the classification of frogs (Sahidullah and Saha, 2012). Meanwhile, MFCCs are not suitable for dealing with recordings with a low signal to noise ratio (SNR). In those previous studies (Huang et al., 2009; Han et al., 2011; Chen et al., 2012; Bedoya et al., 2014) most features used are either based on Fourier transform or transplanted from speech, speaker and music fields. To further improve the recognition and classification performance, it is necessary to develop more accurate species identification methods.

Wavelet analysis has been widely employed for acoustic data, because it can preserve both frequency and temporal information (Ren et al., 2008). Yen and Fu (2002) introduced wavelet packet transform (WPT) for individual frog identification. After applying WPT to the frog calls, energy of all the node coefficient were calculated as features. Then, Fisher's criterion (Yen and Lin, 2000) was used for dimension reduction. Finally, the feature vector after dimension reduction was fed into a neural network classifier for identification. Colonna et al. (2012) proposed to use discrete wavelet transform (DWT) for frog call classification. Based on the node coefficients of DWT, energy, power, zero-crossing rate and pitch of each node coefficients were calculated. However, applying WPT and DWT without any modifications cannot provide a good frequency domain resolution for classifying frog calls.

In this study, the WPD is applied to the frog calls with an adaptive frequency scale for feature extraction. Frog species that are genetically similar often share close advertisement calls (Gingras and Fitch, 2013). Therefore, the dominant frequency which is directly calculated from the trace of advertisement call is an important feature for differentiating frog species. We use dominant frequency to produce the frequency scale for WPD, which is different from using minimum and maximum frequency to generate the frequency scale for WPD in Ren et al. (2008). Specifically, continuous acoustic data are first segmented into syllables using Härmä's method (Harma, 2003). Then, spectral peak tracks are extracted from each syllable where possible. Three features are extracted from each track: track duration, dominant frequency and oscillation rate. Next, a k-means clustering algorithm is applied to the dominant frequency, and the centroids of clustering results are used to generate the frequency scale for WPD. After applying the adaptive frequency scaled WPD to the frog calls, a new feature set named *adaptive frequency scaled wavelet packet decomposition sub-band cepstral coefficients* (AWSCCs) is extracted. Finally, two classifiers, a k-NN classifier and a SVM classifier, are employed for the classification with the proposed feature set.

2. Methods

The architecture of the proposed classification method consists of five modules: syllable segmentation, syllable feature extraction, adaptive frequency scale generation, WPD feature extraction and classification (see Fig. 1). Each module is described in the following sections.

2.1. Sound recording and pre-processing

Two datasets obtained from a commercial recording (Stewart, 1999) and James Cook University (JCU) were selected for this study. Recordings, which were collected from the CD, are two-channel, sampled at 44.10 kHz and saved in MP3 format. All recordings were obtained with a directional microphone and have a high signal to noise ratio (SNR). Each recording includes one frog species, and has a duration ranging from twenty-one to fifty-four seconds. The calls of eighteen frog species recorded in Queensland, Australia were used to develop the detailed methodology. To reduce the subsequent computational burden, all recordings were re-sampled at 16 kHz per second, mixed to mono, and saved in WAV format.

The JCU recordings were obtained from Kiyomi dam (S 19°22′16.0′,E146°27′31.3″) BG creek dam (S19°27′1.23″,E146°24′5.65″) and Stony creek dam (S 19°24′07.0″,E146°2551.3) in Townsville, using Song Meter (SM2) (Xie, 2016). The recordings were stored on 16 GB SD cards in 64 kbps MP3 mono format and have a low

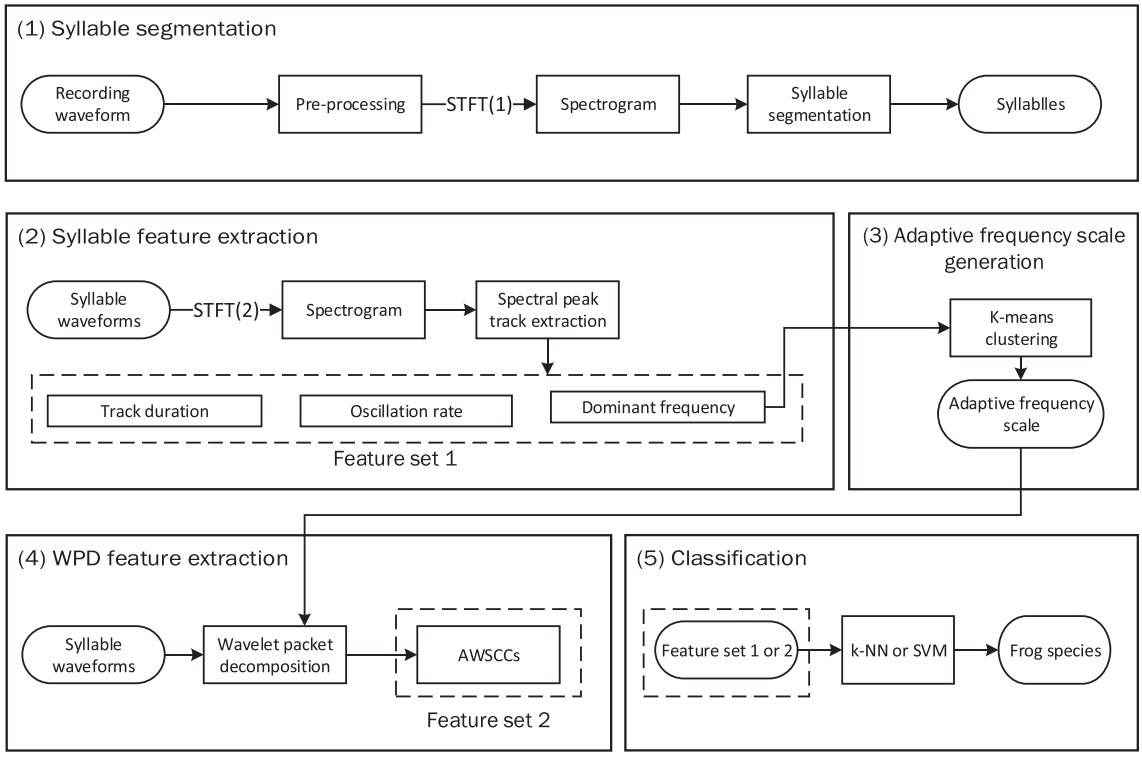

Fig. 1. Block diagram of the frog call classification system. The line of dashes indicates the extracted feature set. AWSCCs is the abbreviation of *adaptive wavelet packet decomposition sub-band cepstral coefficients*. STFT is short-time Fourier transform. For STFT(1), the window function, size and overlap are Kaiser window, 512 samples and 25%. For STFT(2), the window function, size and overlap are Hamming window, 128 samples and 90%. In this diagram, two feature sets are extracted, the description of other feature sets is shown in Fig. 6.