Expert Systems with Applications 42 (2015) 7367–7374

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An incremental technique for real-time bioacoustic signal segmentation

Juan Gabriel Colonna *, Marco Cristo, Mario Salvatierra Júnior, Eduardo Freire Nakamura

Av. Rodrigo Otavio 6200, Institute of Computing (Icomp), Federal University of Amazonas (UFAM), Manaus, Brazil

ARTICLE INFO

Article history:
Available online 28 May 2015

Keywords:
Bioacoustic signal segmentation
Wireless Sensor Networks
Unsupervised learning
Stream data mining

ABSTRACT

A bioacoustical animal recognition system is composed of two parts: (1) the segmenter, responsible for detecting syllables (animal vocalization) in the audio; and (2) the classifier, which determines the species/animal whose the syllables belong to. In this work, we first present a novel technique for automatic segmentation of anuran calls in real time; then, we present a method to assess the performance of the whole system. The proposed segmentation method performs an unsupervised binary classification of time series (audio) that incrementally computes two exponentially-weighted features (Energy and Zero Crossing Rate). In our proposal, classical sliding temporal windows are replaced with counters that give higher weights to new data, allowing us to distinguish between a syllable and ambient noise (considered as silences). Compared to sliding-window approaches, the associated memory cost of our proposal is lower, and processing speed is higher. Our evaluation of the segmentation component considers three metrics: (1) the Matthews Correlation Coefficient for point-to-point comparison; (2) the WinPR to quantify the precision of boundaries; and (3) the AEER for event-to-event counting. The experiments were carried out in a dataset with 896 syllables of seven different species of anurans. To evaluate the whole system, we derived four equations that helps understand the impact that the precision and recall of the segmentation component has on the classification task. Finally, our experiments show a segmentation/recognition improvement of 37%, while reducing memory and data communication. Therefore, results suggest that our proposal is suitable for resource-constrained systems, such as Wireless Sensor Networks (WSNs).

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Forest degradation is a worldwide concern. The success of ecosystem preservation depends on our ability to detect ecological stress in early stages. In this context, anurans (frogs and toads) have been used by biologists as an indicator of ecological stress (Carey et al., 2001). However, monitoring anurans on-site, by human experts, may be too expensive or even unfeasible, depending on the size of the target area. Thus, unassisted monitoring strategies can be adopted, such as the detection of anuran calls using sensor networks (Colonna, Cristo, & Nakamura, 2014; Ribas, Colonna, Figueiredo, & Nakamura, 2012). In such strategies, the sound acquisition is performed by the sensors in a non-intrusive way, which allow us to monitor the environment for a long-term period.

To acquire the anuran calls, sensors are equipped with microphones to gather data (Fig. 1). Unlike other type of sensors, the audio acquisition deals with high sampling frequencies, resulting in a lot of data to be processed and transmitted. The set of all sensors, distributed in an area of interest, comprises a Wireless Sensor Network (WSN) (Akyildiz, Su, Sankarasubramaniam, & Cayirci, 2002). The main advantage of this kind of network is the ability the sensors have to collaborate with each other (Nakamura, Loureiro, & Frery, 2007). As a large number of sensors has to be used, their costs have to be low, which leads to the development of simple devices with limited resources (memory, processing, and bandwidth) (Nakamura, Loureiro, Boukerche, & Zomaya, 2014). Given such constraints, we have to cope with many scientific and technological challenges to effectively use WSNs (Khan, Pathan, & Alrajeh, 2012).

Machine learning techniques with WSNs have already been used for automatic recognition of anuran species (Hu et al., 2009; Potamitis, Ntalampiras, Jahn, & Riede, 2014; Ribas et al., 2012; Wang et al., 2003). These techniques are based on classifiers (e.g. SVM, C4.5 decision trees and kNN) to automate the task of recognizing smaller portions of anuran calls, called syllables. Before classifying the syllables, the calls need to be segmented, i.e., we need to identify the start and the end of every syllable. The precision of the segmentation technique affects the further steps in the species' identification method (Fig. 1), therefore, impacting on the classification performance.

* Corresponding author.
E-mail addresses: juancolonna@icomp.ufam.edu.br (J.G. Colonna), marco.cristo@icomp.ufam.edu.br (M. Cristo), mario@icomp.ufam.edu.br (M. Salvatierra Júnior), nakamura@icomp.ufam.edu.br (E.F. Nakamura).

http://dx.doi.org/10.1016/j.eswa.2015.05.030
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

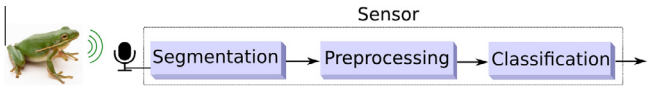

Fig. 1. The three basic steps of a species identification framework.

Fig. 2 shows three syllables of different species recorded under distinct noise conditions. The last segment shows a combination of two syllables. This figure is useful to illustrate the challenge related to proposing a segmentation technique to recognize the four different patterns without compromising the accuracy of the entire recognition system. The challenge of finding the boundaries (*a, b, c, d, e, f, g, h*) can be viewed as an unsupervised binary classification problem, which must be identified when the signal behavior changes. After that, the classification component determines the species whose that segment (syllable) belongs to.

Our main contribution is a real-time incremental technique for segmenting anuran syllables in audio streams.

In contrast to the use of sliding windows (Jaafar, Ramli, & Shahrudin, 2013; Jaafar & Ramli, 2013; Rahman & Bhuiyan, 2012), our incremental strategy stores only simple time series statistics. As such, it has a memory reduction rate of $1/N$ (N is the size of the sliding window), while reducing the algorithmic time complexity from $O(N \times (n/(N-m)))$ to $O(n)$ (N is the window size, m is the window overlapping size, and n is the size of the stream).

As an additional contribution, we derive a methodology for assessing the whole system performance by considering it a multi-level classifier.

The remainder of this work is organized as follows. In Sections 2 and 3, we present the motivation and the problem statement, respectively. Section 4 presents an overview of related work. Our proposal for incremental transformation is presented in Section 6. The evaluation metrics are discussed in Section 7. The parameters, experimental protocol and results obtained are described in Section 8. Finally, in Section 9, we present our conclusions and point out future directions.

2. Motivation

Signal segmentation models have been extensively studied in human speech recognition. However, these models are not well suited for anuran calls, which have different characteristics (Rickwood & Taylor, 2008). For a better feature extraction and improvements of species classification, it is important to select the most representative parts of an anuran call, since these calls usually contain long periods of environmental noise (Evangelista, Priolli, Silla, Angelico, & Kaestner, 2014).

The majority of the approaches for automatic call segmentation involves non-sequential procedures that consume large amounts of memory (Garcia, Marcias-Toro, Vargas-Bonilla, Daza, & López, 2014; Härmä, 2003; Xie et al., 2015). Moreover, these types of approaches are not suitable for data stream scenarios, in which large amounts of data must be processed in real-time by resource-constrained systems/networks (Nakamura et al., 2014). As segmentation is the first step of the recognition framework (Fig. 1), this has a direct impact on the species identification rate. Therefore, we must understand the relationship between the segmentation and the classification components.

3. Problem statement

A syllable is one elementary bioacoustic unit for classification. A continuous call, emitted by an individual frog, is composed of several syllables repeated along the time. Fig. 3 shows a typical call of the *Leptodactylus hylaedactylus* species with three syllables. The beginning, middle, and end-points of the syllables are delineated by vertical lines depicting three different types of changes that characterize the vocalization: (1) an abrupt change in the signal level – e.g., the change from noise to syllable indicated by the first vertical dotted line; (2) a gradual change, upward or downward – e.g., a gradual increase of noise or a soft signal attenuation, as seen between the second and third vertical dotted lines; and (3) recurrent change patterns – e.g. the three similar syllables repeated over time.

The problem of syllable segmentation is to detect the beginning and the end of a syllable. Thus, considering a specific audio stream, we aim at extracting all syllables. The time intervals between the syllables (ambient sound/noise) are not useful to detect the animal. In fact, the noise sections are discarded, because: (a) they increase the misclassification rate; (b) they increase transmission costs; and (c) they reduce the WSN lifetime. For this reason, in real situations, it is convenient to detect changes in the monitored signals to decide when to start and stop data communication or data processing. How to measure the quality of the segmentation is an additional problem, related to assign the correct species to the corresponding extracted syllables (classification).

4. Related work

Automatic sound segmentation has been widely studied, usually focusing on music and human voice streams (Theodorou, Mporas, & Fakotakis, 2014). In such studies, it is common to prioritize robust solutions even if it results in higher costs. Foote (2000) focused on music using a kernel function for segmentation, while Sarkar and Sreenivas (2005) addressed speech, employing the Average Level Crossing Rate (ALCR). The problem of segmenting different sources like music, speech, and environmental sound from movies was addressed by Giannakopoulos, Pikrakis, and Theodoridis (2008) by using a technique based on eight spectral band frequencies.

Similar expensive approaches were also employed in the study of bioacoustic signals. For instance, Potamitis et al. (2014) applied the costly Hartley Transform, with good results. Evangelista et al. (2014) used the spectrogram to extract two features that represent the syllables. To obtain these syllables, the histogram of the energy

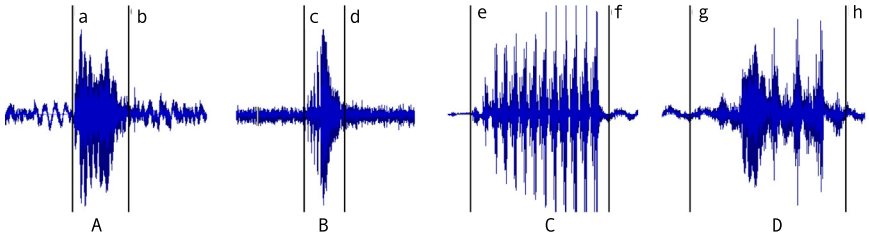

Fig. 2. Four syllable patterns under different noise conditions.

Download English Version:

https://daneshyari.com/en/article/10322190

Download Persian Version:

https://daneshyari.com/article/10322190

Daneshyari.com