Ecological Informatics 25 (2015) 14–21

Contents lists available at ScienceDirect

Ecological Informatics

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

Decision support for the efficient annotation of bioacoustic events

Anthony Truskinger *, Michael Towsey, Paul Roe

QUT Bioacoustics, Science and Engineering Faculty, Queensland University of Technology, Brisbane, Australia

ARTICLE INFO

Article history:
Received 14 July 2014
Received in revised form 24 September 2014
Accepted 3 October 2014
Available online xxxx

Keywords:
Similarity search
Bioacoustics
Annotations
Semi-automated
Decision support
Faunal vocalisation

ABSTRACT

Acoustic sensors allow scientists to scale environmental monitoring over large spatiotemporal scales. The faunal vocalisations captured by these sensors can answer ecological questions, however, identifying these vocalisations within recorded audio is difficult: automatic recognition is currently intractable and manual recognition is slow and error prone. In this paper, a semi-automated approach to call recognition is presented. An automated decision support tool is tested that assists users in the manual annotation process. The respective strengths of human and computer analysis are used to complement one another. The tool recommends the species of an unknown vocalisation and thereby minimises the need for the memorization of a large corpus of vocalisations. In the case of a folksonomic tagging system, recommending species tags also minimises the proliferation of redundant tag categories.

We describe two algorithms: (1) a "naïve" decision support tool (16%–64% sensitivity) with efficiency of $O(n)$ but which becomes unscalable as more data is added and (2) a scalable alternative with 48% sensitivity and an efficiency of $O(\log n)$. The improved algorithm was also tested in a HTML-based annotation prototype. The result of this work is a decision support tool for annotating faunal acoustic events that may be utilised by other bioacoustics projects.

1. Introduction

Acoustic sensors are an effective method for the large scale monitoring of fauna within an ecosystem. They can objectively record data over large spatiotemporal scales and the recordings can be used for ecological tasks such as determining species presence/absence. However, raw audio data is opaque — it must be analysed before it is of any use. Manual processing of audio (e.g. by having an appropriately qualified expert listen to recordings) can identify species accurately but is slow. On average, it required 2 min of listening for an expert to identify the bird species in 1 min of audio (Wimmer et al., 2013b). On the other hand, automated methods, although they hold out the promise of being fast, do not have the accuracy currently required for ecological studies (Potamitis et al., 2014). There has been some success with various single-species recognisers (Brandes et al., 2006; Hu et al., 2009; Kasten et al., 2010; Towsey et al., 2012; Wimmer et al., 2013a) and some multi-species recognisers (Acevedo et al., 2009; Anderson et al., 1996; Harma, 2003). A state-of-the-art recogniser has been reported by Stowell and Plumbley (2014) but its application to the pre-prepared *lifeclef2014* dataset (Joly et al., 2014) does not necessarily translate to the rigorous requirements of an ecological study. Even capable automatic recognisers often require manual verification (Wimmer et al., 2013a).

An alternative analytical approach for recordings of faunal vocalisations is to extract ecological indices which point to the presence of animal vocalisations of interest rather than identifying the actual species (Bart, 2005; Depraetere et al., in press; Gage et al., 2001; Gasc et al., 2013; Pieretti et al., 2011; Towsey et al., 2014). This approach is part of the emerging field of *soundscape ecology* that views the acoustic world from an ecological perspective rather than a species perspective (Pijanowski et al., 2011).

Humans can become excellent classifiers of bioacoustic events given sufficient training but manual analysis of audio data is a laborious process, the more so for experts. It is also expensive. However, humans can work more efficiently if given appropriate technical support. This so-called *semi-automated* approach combines the complementary strengths of human and computer. In this paper, we explore a *semi-automated* approach to the identification of animal vocalisations, primarily but not exclusively due to birds. When annotating, users are given a short sample of audio and its pictorial representation as a spectrogram; the user is required to identify the species making the call. Decision support takes the form of a "suggestion tool" that shows similar *labelled* samples of audio and spectrograms to the user. We have previously reported a proof-of-concept decision support system embedded in a website (Truskinger et al., 2011). The purpose of that paper was to test the effectiveness of the suggestion tool on the performance of a mix of expert and non-expert participants. The authors report a slight (but statistically significant) increase in the participant classification rate but not in their classification accuracy. Interviews with participants indicated that the suggestion tool was potentially helpful but needed to

* Corresponding author at: S Block, Level 10, S1002, Gardens Point Campus, 2 George St, Brisbane, QLD 4000, Australia. Tel.: +61 7 3138 9381.
E-mail address: anthony.truskinger@student.qut.edu.au (A. Truskinger).

http://dx.doi.org/10.1016/j.ecoinf.2014.10.001
1574-9541/

be more accurate. The participant feedback provides the motivation for the work described in this paper.

The suggestion tool reported by Truskinger et al. (2011) relied on 400 *reference annotations*. A reference annotation is one determined by experts as being a good exemplar of its class. In this paper, we investigate the hypothesis that a decision-support system dependent on typical annotations (as opposed to exemplars) would improve in accuracy. The remainder of this paper is organised as follows: Section 2 describes related work. Sections 3, 4 and 5 describe our methodology, results, and discussion respectively. The final sections describe future work and conclude.

2. Related work

Annotating multimedia data with tags is a common practice on the web. Examples of multimedia annotation include: Flickr (images), SoundCloud (sound), YouTube (video formats), and Vannotea (Schroeter et al., 2006) which can annotate most multimedia formats. This research focuses on annotating audio data for ecological science. Similar research projects have cultivated libraries of audio recordings that have been labelled (usually the entire recording is labelled). The 'Jacques Vielliard' dataset maintained by UNICAMP (Cugler et al., 2011) and the Berlin Sound Archive (Bardeli, 2009) are two examples. These libraries are excellent resources; however, the majority of their recordings are not acoustic sensor recordings. Instead, they are usually targeted and have high *signal-to-noise ratios* (SNRs).

Analysts, including novices, find the detection and isolation of bioacoustic events from background events to be easier than the classification of those events. Because a large corpus of audio patterns must be memorised in order to classify events, few people have enough experience or skill to identify all faunal vocalisations by recall alone. Even a geographically constrained set of recordings from just one site can contain hundreds of vocalising species. Some of these species, especially birds, have more than one form of vocalisation. For example, at QUT's SERF facility, located in the Samford Valley, Queensland, Australia, 460 unique tags (*classes*) have been applied to 100 species, found in 80000 bioacoustic events from six days of data (Wimmer et al., 2013a).

Some experts can aurally classify large numbers of bird species by recall alone. These experts have had many years of training as ornithologists or through recreational *birding* activities. However, their memorised knowledge is limited to the geographical areas where they have had experience; different environments often mean different sets of species. Vocalisations of species can also vary between regions creating further difficulty (Kirschel et al., 2009).

Nevertheless, humans are exceptional at pattern recognition tasks (Sroka and Braida, 2005) and identification of acoustic events becomes easier when a spectrogram accompanies the audio data (Wimmer et al., 2013a). Most analysts can discern visual differences between spectrogram features with ease. Human selected discriminating features are creative, often qualitative, and describe aspects of an object that are hard to quantify (Feyyad, 1996). Humans can discriminate audio-patterns even in noisy, degraded, or overlapping signals (Rusu and Govindaraju, 2004). To summarize, any method to augment the skills of human analysts should utilise their exceptional comparison skills and place less emphasis on recall of prior knowledge.

A decision support tool for bioacoustic events imposes a set of constraints on the *user interface* (UI). The autocomplete box, a similar but far less complex UI mechanism, suggests possible textual matches within milliseconds, sometimes from remote sources. Likewise, an effective decision support tool must also provide results in sub-second times as its utility depends on its response-time. The recommended response-time for page navigation is sub-second and for interactive visual components is less (Miller, 1968; Nielsen, 1999).

The task of matching a 'sound-bite' to a larger database of audio for the purpose of classification has been previously accomplished in both the ecological acoustics and music fields (Bardeli, 2009; Kasten et al., 2012; Wang, 2006). Because vocalisations occur in noisy environments and vary greatly by region, music matching methods are ineffective for matching faunal vocalisations (Cugler et al., 2011). Currently, there is no effective system for automated content-based similarity search of faunal vocalisations. The existing partial-solutions to similarity search all require signal processing to extract features and complex classification algorithms. Given the immense volume of data collected by acoustic-sensors, the difficulty of the classification task, and the need to generate suggestions quickly, the suggestion task lends itself to a metadata-based solution.

Other sound ecology software packages have been created that may benefit from the approaches in this paper. The Pumilio project is an open source software package that allows researchers to store audio recordings (Villanueva-Rivera and Pijanowski, 2012). Pumilio allows recordings to be uploaded, analysed, and tagged with metadata, through a web interface. Similarly, The REAL digital library is an archive of sensor recordings accessible through a web interface. The REAL project also allows automated analysis and has search capabilities (Kasten et al., 2012).

3. Experimental method

Increasing the quantity of training data is a standard approach used to increase the accuracy of supervised machine-learning problems (Zhu et al., 2012). We have previously published results for an experiment where the training data consisted of 400 exemplar annotations; that is, the canonical or best examples of calls for each class. However, most ordinary acoustic events in real recordings of the environment are distorted by noise or overlapping events. Furthermore, the majority of recorded vocalisations have low signal-to-noise ratios. Low SNR is seen as an effect of the combination of the inverse-square law and the probable distribution of fauna around a sensor; it is likely that more vocalising individuals will be further from the microphone. In this work, we investigate the hypothesis that increasing the proportion of poorer quality calls (relative to high SNR canonical calls) within the training data will increase the accuracy of the resulting decision support tool.

A large increase in the quantity of training data affects the choice of algorithmic approach (Deng et al., 2010). For the decision support tool, new algorithms are tested for their scalability and ease of implementation. To achieve scalability, the feature set was kept to a minimum. In particular, we focused on easy-to-extract features derived from the meta-data of an annotated call as opposed to audio-content features.

The experimental framework for this research was to evaluate performance for multiple simulations of the decision support tool over different combinations of datasets, algorithmic components, and feature sets. This section describes the components of the simulations.

3.1. Datasets

Two datasets were used for the experiment: the *Full* dataset and the *Reference* dataset. Both datasets use the same testing data. Table 1 has a summary breakdown on the number of annotations and their tags, for each dataset.

The *Full* dataset consists of annotations generated by human analysts, in audio recordings taken from the QUT Samford Ecological Reserve Facility (SERF), located north-west of Brisbane, Queensland, Australia. The annotated dataset was produced by Wimmer et al. (2013a). The vegetation at SERF is mainly open-forest to woodland comprised primarily of *Eucalyptus tereticornis*, *Eucalyptus crebra* and *Melaleuca quinquenervia* in moist drainage. There are also small areas of gallery rainforest with *Waterhousea floribunda* predominantly fringing the Samford Creek to the west of the property, and areas of open pasture along the southern border. Faunal vocalisations were analysed by experts producing 473 call types (tags) for 96 species across four sites. The majority of the species identified were *Aves*; however, there are examples of crickets, frogs, and marsupials in the dataset. The most frequently detected species include the Rufous Whistler