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).

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.

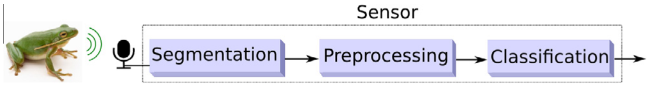

Fig. 1. The three basic steps of a species identification framework.

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

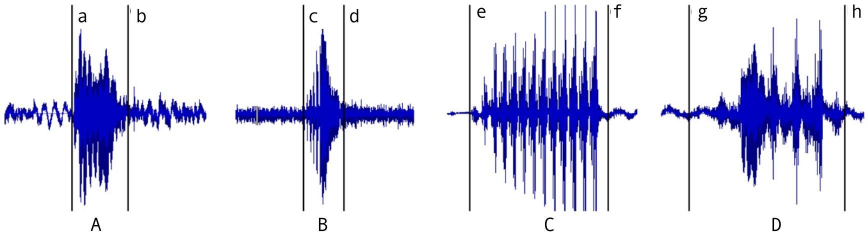

Fig. 2. Four syllable patterns under different noise conditions.

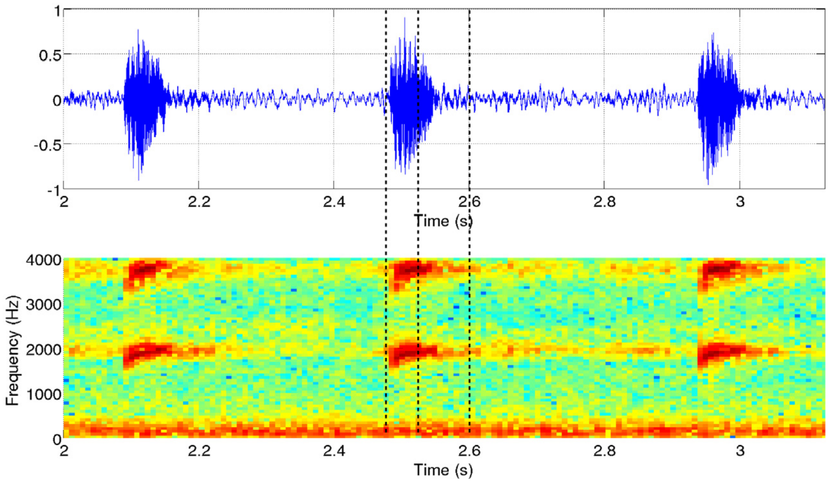

Fig. 3. Three syllables of the *Leptodactylus hylaedactylus* call: signal amplitude (above) and spectrogram (below).

of the signal and its spectral centroid are computed, thereafter a median filter is applied to find the threshold levels and perform the segmentation. The quality of the segmentation depends on the threshold values and, to properly choose these values, the knowledge of the entire signal is required.

Frequency based methods have been shown to be more robust to background noises. For example, Härmä (2003) and Xie et al. (2015) apply image processing techniques to the spectrogram of the signal (Fig. 3). These approaches repeat several processing steps, beginning from the best syllable until extract all syllables. This procedure looses the temporal order, so it is inadequate for real time scenarios. Beyond the computational cost of the Fourier Transform (FFT), more memory is required to handle the entire image, which makes this approach unsuitable for resource-constrained systems.

Neal, Briggs, Raich, and Fern (2011) use a Random Forest classifier to recognize spectrogram frames. The use of a supervised classifier has two advantages: (1) it is robust to retrieval the desired syllables and (2) segmentation and classification may be performed in a single step. However, this approach is computationally expensive, because of the high cost associated with FFT and the use of many decision tree models. Furthermore, the set of possible animals must be known a priori.

A naive strategy to segment audio samples consists in finding threshold values, such that if the sample amplitude is smaller than the threshold, the sample is classified as noise (Cheng, Sun, & Ji, 2010; Colonna, Ribas, dos Santos, & Nakamure, 2012; Huang, Yang, Yang, & Chen, 2009). While simple, such an approach is very sensitive to random noise of short duration and high amplitude. To improve segmentation accuracy, Fagerlund (2007) used the energy value of the signal in dB as a temporal feature to make the method more robust to noise. This approach has the following shortcomings: (1) devices need enough memory to store the whole stream being processed; (2) the whole audio stream is processed several times to segment the existing syllables. As a consequence, the syllables cannot be segmented by resource-constrained devices in real time.

Other methods are based on more restricted strategies that use a limited number of successive samples from a sliding window (Jaafar et al., 2013; Jaafar & Ramli, 2013; Rahman & Bhuiyan, 2012). In particular, these methods used a combination of two low-cost temporal features to improve the segmentation performance. These features are the signal Energy (E) and the Zero Crossing Rate (ZCR). For every iteration of this method, a new window is processed sequentially. This method is more resilient to noise, because it requires that the two thresholds are satisfied at the same time, one for Energy and one for ZCR.

An indirect way to extract the fundamental frequency of the anuran calls, without performing the FFT, is to compute the pitch based on the auto-correlation function. This feature, used to find strong frequency components in the signal, was employed by Garcia et al. (2014). The authors conducted an interesting evaluation, which compares the totally recovered syllables to the manually segmented syllables. This evaluation approach is more consistent with a segmentation than a classification task. Unfortunately, the authors have not reported the rate of missing syllables.

Most of the previous segmentation methods are not suitable to real-time processing in resource-constrained systems, because they use large segments of the sound signal. These methods are not appropriate in our context, since we consider that the future of the incoming signal is unknown. Table 1 summarizes the key features of each related work. We also observed that the use of temporal features along with sequential processing result in simpler implementations that require fewer hardware resources. Furthermore, window-based methods are able to achieve good performance by remembering a fixed number of past samples, which suggests that past information can be incrementally updated.

Another contribution of our work is related to how we can assess a sound classification system that performs segmentation. In a previous work, Han, Muniandy, and Dayou (2011) have studied the impact of the segmentation on the classification rate. As observed by the authors, there is no consensus on how to asses the final performance, because of issues such as misclassifications due to unrecognized segments. Thus, to evaluate our framework, we propose a new way to evaluate the whole system, described in Section 7.2.

5. Background

The signal energy allows us to know when the amplitude increases, while the ZCR provides an approximation of the main frequency. These two temporal features, commonly used in audio precessing tools, are given by:

$$E_N = \frac{1}{N}\sum_{i=1}^{N} x_i^2, \tag{1}$$

Table 1
Summary of related works. The abbreviations Acc, P, R, F1, and AEER stands for Accuracy, Precision, Recall, F-Score, and Acoustic Event Error Rate, respectively.

Authors	Features	Procedure	Evaluation form
Han et al. (2011)	Spectral-entropy	Manual	Classification Acc.
Rahman and Bhuiyan (2012)	E and ZCR	Non-sequential	No. of syllables
Jaafar et al. (2013)	E and ZCR	Sequential	No. of syllables Classification Acc.
Huang et al. (2009)	Amplitude	Non-sequential	Classification Acc.
Colonna et al. (2012)	Amplitude	Non-sequential	Classification Acc.
Cheng et al. (2010)	Amplitude	Non-sequential	No. of syllables Classification Acc.
Fagerlund (2007)	Energy	Non-sequential	Classification Acc.
Neal et al. (2011)	Spectrogram	Sequential	ROC curves
Härmä (2003)	Spectrogram	Non-sequential	No. of Syllables Confusion Matrix
Xie et al. (2015)	Spectrogram	Non-sequential	No. of syllables Classification Acc.
Evangelista et al. (2014)	Energy Spectral Centroid	Non-sequential	Classification Acc.
Potamitis et al. (2014)	Hartley Transform	Sequential	Classifier P, R and F1.
Our Approach	Incremental E Incremental ZCR	Sequential Real-Time	AEER. Whole system P, R and F1.

$$ZCR_N = \frac{1}{2N}\sum_{i=1}^{N}|\text{sign}(x_i) - \text{sign}(x_{i-1})|, \tag{2}$$

in which x_i is the amplitude of the audio signal and N the frame size. The function sign() is defined as:

$$\text{sign}(x_i) = \begin{cases} +1, & \text{if } x_i \geqslant 0 \\ -1, & \text{if } x_i < 0 \end{cases} \tag{3}$$

By applying a sliding window, a set of consecutive frames is generated. If N is much longer than the size of the syllable, many unnecessary values will be used to compute E_N and ZCR_N, probably causing syllable losses. In the opposite case, when N has a very short duration, impulse noises can be interpreted as part of a syllable resulting in a false detection. The requirement that two thresholds (T_E for E_N and T_Z for ZCR_N) must be satisfied simultaneously may prevent this issue (Jaafar et al., 2013; Jaafar & Ramli, 2013; Rahman & Bhuiyan, 2012). Thus, a combination rule is defined as:

$$T_h = \begin{cases} 1, & \text{if } E_N \geqslant T_E \text{ and } Z_N \geqslant T_Z \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The memory cost associated with this approach is equal to N and the time to process an entire signal, of size n, depends on the overlap parameter m ($O(N \times (n/(N-m)))$). Note that parameters N and m affect the segmentation accuracy (Section 8), as it is hard to define a pair of values that operate correctly through successive windows.

Herein, from now on, we refer to this method as EZ-WIN.

6. A new incremental segmentation approach

In this section, we present an incremental variation of segmentation techniques for extracting anuran calls from audio streams. In contrast to sliding-window approaches, the incremental proposal is more efficient regarding memory and processing usage.

According to Jaber (2013), incremental techniques should satisfy the following requirements: (a) adaptability – to adapt to gradual changes without ignoring the abrupt ones; (b) knowledge transfer – to deal with new samples without forgetting past decisions; (c) appropriate reaction time – to detect change as fast as possible; and (d) low error rate – to avoid the increase of false positive and false negative rates. Based on these requirements, we are interested in segmentation techniques that are sensitive to change, robust to noise and false alarms, and efficient regarding memory, processing time, and energy. To accomplish this, we redesign previous methods (Jaafar & Ramli, 2013) to remember past information by using metrics that can be calculated incrementally, without relying on data windows.

We start by presenting an incremental version for computing the Energy Eq. (1). This equation averages all observed x_i^2 values in a window of size N. We can rewrite the Energy E_n with weights $w_1, w_2, \ldots w_n \geqslant 0$, similarly to the exponentially-weighted mean (Finch, 2009) $E_n = \frac{1}{W_n}\sum_{i=1}^{n} w_i x_i^2$, in which $W_n = \sum_{i=1}^{n} w_i$ and $\alpha = \frac{w_n}{W_n}$. Thus:

$$E_n = \frac{1}{W_n}\left(w_n x_n^2 + \sum_{i=1}^{n-1} w_i x_i^2\right)$$

$$E_n = \frac{1}{W_n}(w_n x_n^2 + W_{n-1}E_{n-1})$$

$$E_n = \frac{1}{W_n}(w_n x_n^2 + (W_n - w_n)E_{n-1})$$

$$E_n = \frac{1}{W_n}(W_n E_{n-1} + w_n(x_n^2 - E_{n-1}))$$

$$E_n = E_{n-1} + \frac{w_n}{W_n}(x_n^2 - E_{n-1})$$

$$E_n = E_{n-1} + \alpha(x_n^2 - E_{n-1}) \tag{5}$$

Eq. (5) does not require a window of size N to estimate current energy E_n. In this new formulation, the parameter α controls the trade-off between the weight of the current sample and the historical data.

Similarly, we can derive an incremental version of ZCR as $Z_n = \frac{1}{2W_n}\sum_{i=1}^{n} w_i y_i$, where $y_i = |sign(x_i) - sign(x_{i-1})|$. Thus, we can rewrite ZCR (Eq. (2)) as:

$$2Z_n = \frac{1}{W_n}\left(w_n y_n + \sum_{i=1}^{n-1} w_i y_i\right)$$

$$2Z_n = \frac{1}{W_n}(w_n y_n + 2W_{n-1}Z_{n-1})$$

$$2Z_n = \frac{1}{W_n}(2W_n Z_{n-1} + w_n(y_n - 2Z_{n-1}))$$

$$2Z_n = 2Z_{n-1} + \frac{w_n}{W_n}(y_n - 2Z_{n-1})$$

$$2Z_n = 2Z_{n-1} + \alpha(y_n - 2Z_{n-1}),$$

which, divided by 2, results in Eq. (6).

$$Z_n = Z_{n-1} + \alpha\left(\frac{|sign(x_n) - sign(x_{n-1})|}{2} - Z_{n-1}\right). \tag{6}$$

Eqs. (5) and (6) provide E and ZCR values for each input sample without sliding windows. These equations are able to adapt to gradual changes, such as the progressive increase of noise without losing abrupt changes. They support knowledge transferring since they do not forget past decisions when new samples are processed. Both equations have few parameters to adjust and satisfy memory and processing constraints of the sensor nodes. The memory cost in the case of equations is $O(1)$ and the time cost is $O(n)$. In this work, we refer to this incremental method as EZ-I.

To reduce the probability of false positives, we can use a mode filter.[1] We refer to the version of EZ-I with a mode filter as EZ-IMF. This modification allows the current sample to be identified as a syllable if most of the past S samples were syllables. Algorithm 1 presents EZ-IMF and Fig. 4 illustrates its application. In this algorithm, E_n and Z_n are updated for each sample x_n (lines 2–3). If both thresholds are overstepped and the mode filter, implemented with a single mc counter, is smaller than S, then mc is incremented by one (lines 4 and 5). If the sample is not identified as a syllable, the mc is decremented (lines 6 and 7). Once identified as a syllable, the sample is sent for further processing if it satisfies the condition of the mode filter (line 10).

Algorithm 1. EZ-IMF Segmenter.

```
 1: function ChangeDetection $x_n, \alpha, E_{(n-1)}, Z_{(n-1)}, S$
 2:     $E_n = E_{(n-1)} + \alpha(x_n^2 - E_{(n-1)})$
 3:     $Z_n = Z_{n-1} + \alpha\left(\frac{|sign(x_n) - sign(x_{n-1})|}{2} - Z_{n-1}\right)$
 4:     if $E_n \geqslant T_E$ and $Z_n \geqslant T_Z$ and $mc < S$ then
 5:         mc = mc + 1
 6:     else if $mc > 0$ then
 7:         mc = mc−1
 8:     end if
 9:     if mc $\geqslant \frac{S}{2}$ then
10:         Send $x_n$
11:     end if
12: end function
```

Because of the α parameter, our approach has a slight delay to detect the beginning of each syllable. The same problem is also observed in the methods based on sliding windows, because of the window length and the overlapping factor.

7. Evaluation methodology

To evaluate our method, we compare the segmentation result against the perfect segmentation (manual segmentation by a specialist), quantifying three type of errors: point-to-point, boundaries-to-boundaries and event-to-event.

To count event-to-event errors, we use a metric that was designed for segmentation algorithms, namely, the Acoustic Event Error Rate (AEER). The AEER is frequently used in context detection (Giannoulis et al., 2013), and it is defined as:

$$\text{AEER} = \frac{\text{D} + \text{I} + \text{S}}{\text{N}}, \tag{7}$$

in which N is the number of syllables in each audio clip, D is the number of missed syllables, I the number of extra syllables, and S the number of replaced syllables computed[2] as $S = min(D, I)$. This metric considers that an event is correctly segmented if it starts and ends within ±100 ms of the event's real boundaries and if it has at least 50% of the real syllable timespan. In addition, duplicated events are considered false alarms. Thus, in the best case, $\text{AEER} = 0$.

To measure the accuracy of the estimated boundaries, compared to the perfect segmentation (boundaries-to-boundaries), we propose to apply the approach developed by Scaiano and Inkpen (2012), known as WinPR. WinPR counts errors from boundaries so that close errors are not masked as in AEER. Once true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are defined, traditional metrics such as precision, recall, and F-Score (cf. Eqs. (8)–(10)) can be calculated. These metrics are very useful for comparing the retrieved signal points that are relevant and the fraction of relevant points that are retrieved. The higher the value of these metrics, better is the result of the boundaries. These metrics are defined as:

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{8}$$

$$\text{R} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{9}$$

$$\text{F1} = \frac{2 \cdot \text{P} \cdot \text{R}}{\text{P} + \text{R}}. \tag{10}$$

Often methods for audio detection and segmentation are evaluated by using metrics based on a decision table or confusion matrix. Thus, each point of the real signal is compared to each point of the automatically segmented signal (point-to-point approach). The result of the segmentation is binary (signal or not signal) and TP, TN, FP and FN figures are interpreted slightly different from WinPR. Thus, for completeness, we also present True Positive Rate (TPR), False Negative Rate (FNR) and the Matthews Correlation Coefficient (MCC) (Powers, 2007), according to the point-to-point approach (cf. Eqs. (11)–(13)). Higher values of TPR and lower values of FNR indicate little signal losses. As for MCC, the output values may vary from −1 to 1, indicating higher ($MCC = 1$), lower ($MCC = 0$) and negative ($MCC = -1$) correlation.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{11}$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}}, \tag{12}$$

$$\text{MCC} = \frac{(\text{TP} \cdot \text{TN}) - (\text{FP} \cdot \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \tag{13}$$

As the dataset is not balanced, we provide macro-averaging of the metrics, as recommended by Sokolova and Lapalme (2009). In other words, we first compute the metrics for each species and, then, average of these values.

7.1. Species classifier

To evaluate the whole system, we first segmented the syllables and annotated them. Thus, for each segmented syllable, we identified the corresponding species. To automatically classify the syllables, they are represented by a set of twelve Mel Frequency Cepstral Coefficients (MFCC) obtained from a filter bank with 24 filters. The classifier we use is kNN (with $k = 1$) and classification results were evaluated by using Leave-One-Out Cross-Validation. As our main objective is to show the performance of the complete system, we chose this feature set and classifier based on the framework defined by Colonna et al. (2012). In next section we present our approach for the recognition problem as a multi-level system where two tasks, segmentation and classification, are chained.

7.2. Multi-level classifier

As pointed out before, the segmentation and classification tasks are treated as a multi-level classifier with unsupervised and supervised layers. In this setting the output of the segmenter (the first classifier) is the input for the species recognizer (the second

[1] A mode filter considers the most frequent value that was observed as the correct value.

[2] The implementation of Giannoulis et al. (2013) can be found at http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/.

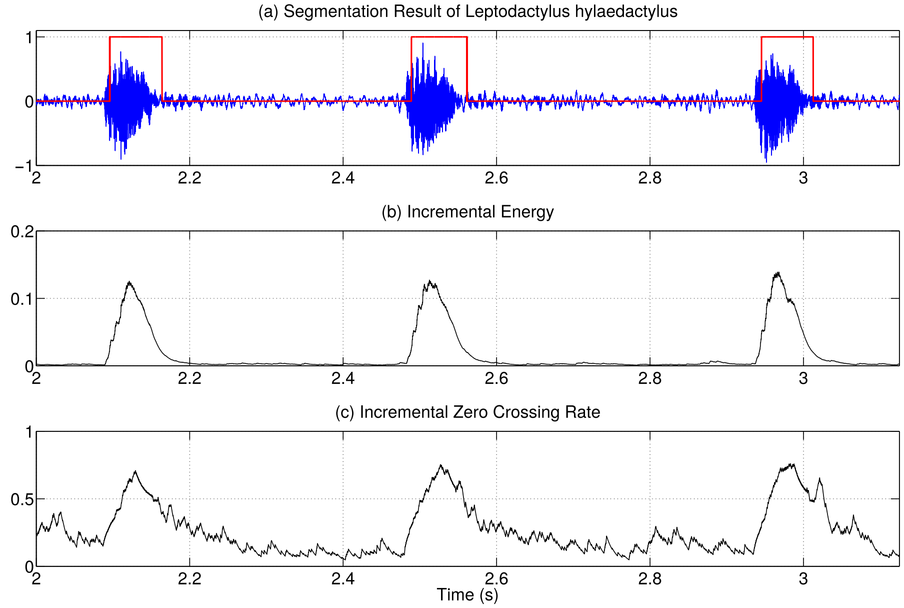

Fig. 4. Three syllables of *Leptodactylus hylaedactylus* call. (a) Segmentation output in red by EZ-IMF. (b) Plot of incremental Energy (Eq. (5)). (c) Plot of incremental ZCR (Eq. (6)). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

classifier) that makes the final decision. To illustrate their operation, suppose that a true syllable is recognized by the segmenter (true positive TP_s). Given that input, the second classifier may produce one of the four possible outputs: “it is the target species” (TP_c), “it is not the target” (TN_c) and two type of errors, FP_c and FN_c. Thus, the true positive rate obtained by the final classifier is given by Eq. (14).

From the previous example, it is clear that, for each different input (signal or noise), the final classification result depends on the segmenter response. All possible combinations (cf. Fig. 5) of correct and incorrect decisions made by segmenter and classifier are summarized by Eqs. (14)–(17):

$$\text{TP}_f = \text{TP}_s \cdot \text{TP}_c, \tag{14}$$

$$\text{TN}_f = \text{TP}_s \cdot \text{TN}_c + \text{TN}_s, \tag{15}$$

$$\text{FN}_f = \text{TP}_s \cdot \text{FN}_c + \text{FN}_s, \tag{16}$$

$$\text{FP}_f = \text{TP}_s \cdot \text{FP}_c + \text{FP}_s(\text{TP}_c + \text{FP}_c + \text{FN}_c + \text{TN}_c), \tag{17}$$

in which subscripts s, c and f stand for the outputs of segmenter, classifier and final result, respectively.

In this system, when the segmenter produces a TN_s or a FN_s, nothing is sent to the classifier, preventing a misclassification or causing a syllable losses. This characteristic leads to less combinations of decisions between segmenter and classifier. In this case, the final result can be computed by Eqs. (15) and (16).

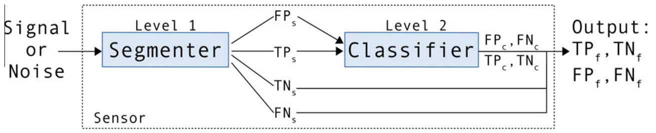

Fig. 5. Interaction between segmenter and classifier as a whole system. Only recognized segments are sent to the species classifier.

Table 2
Average results for EZ-I, EZ-IMF ($S\;=40$) and EZ-WIN (with N varying from 64 to 1024 and overlap = 50%).

Technique	TPR	FNR	MCC	P	R	F1	AEER
EZ-I	0.53	0.46	0.59	0.93	0.53	0.62	19.61
EZ-IMF	0.51	0.48	0.57	0.90	0.51	0.60	4.36
EZ-WIN, N = 64	0.12	0.87	0.24	0.99	0.12	0.19	6.69
EZ-WIN, N = 128	0.14	0.85	0.27	0.99	0.14	0.22	5.66
EZ-WIN, N = 256	0.16	0.83	0.29	0.93	0.16	0.24	3.05
EZ-WIN, N = 512	0.14	0.85	0.27	0.92	0.14	0.21	1.58
EZ-WIN, N = 1024	0.07	0.92	0.14	0.57	0.07	0.11	1.34

8. Experiments and results

To evaluate our methods, we used a dataset with 15 frogs from seven different species for a total of 896 syllables. The audio signals in this dataset have a frequency of 8 kHz and 8 bits per sample. The α value was set to 0.01 and we used the thresholds values $T_E = 0.02$ and $T_Z = 0.3$. The experiments were carried out in Matlab (detailed scripts and the dataset are available at http://bit.ly/1b8bvyE).

Table 2 shows a comparison between our methods (EZ-I and EZ-IMF) and the sliding window versions (EZ-WIN). For EZ-WIN, overlap was set as 50%, a value commonly used. We used window sizes of 64, 128, 256, 512, and 1024 units (the larger the size, the larger the memory used to remember the past). From this table, it is clear that large window sizes improve AEER at the expense of memory. Small window sizes improve precision, but decrease recall and increase AEER. These results show that the recall of the sliding window techniques is consistently weak compared to our approaches.

In addition, our method with the lowest error rate, EZ-IMF, misses about 50% of the segments annotated by the human expert. In general, given the higher recall, the higher TPR and the lower FNR of our methods, it is clear that they have a better trade-off between the number of missing syllables and the transmission of unnecessary samples. Finally, the MCC values of our approaches (EZ-I and EZ-IMF) indicate more positive relationship than EZ-WIN approaches.

When comparing EZ-I and EZ-IMF, the EZ-I performs very poorly regarding AEER. This large error rate occurred because EZ-I generates many short segments for the species *Osteocephalus O.* Most of the segments were correctly recognized as part of

Table 3
Comparison between EZ-IMF and EZ-WIN considering each species.

Species	Syllables	EZ-IMF			EZ-WIN$_{512}$			
		TPR	MCC	AEER	TPR	MCC	AEER	
Leptodactylus H.	453	0.63	0.74	0.01	0.20	0.39	0.71	
Leptodactylus F.	218	0.65	0.71	0.04	0.15	0.33	0.81	
Ameerega T.	88	0.26	0.35	1.24	0.18	0.28	0.96	
Adenomera A.	73	0.51	0.56	0.27	0.24	0.43	1.07	
Hyla Minuta	40	0.43	0.55	1.88	0.10	0.24	1.12	
Rhinella G.	17	0.86	0.72	23.85	0.02	0.05	5.18	
Osteocephalus O.	7	0.27	0.46	7.73	0.00	0.08	2.00	
Average		0.51	0.57	4.36		0.14	0.27	1.58

Table 4
Perfect segmentation (PS), Precision, Recall and F-Score of the entire system. Approximation of memory used for the segmentation (Byte).

Metrics	PS	W$_{64}$	W$_{128}$	W$_{256}$	W$_{512}$	W$_{1024}$	EZ-I	EZ-IMF
Precision	0.99	0.98	**0.99**	**0.99**	**0.99**	0.83	0.92	0.94
Recall	0.96	0.09	0.12	0.15	0.12	0.06	0.42	**0.49**
F1	0.98	0.16	0.21	0.25	0.21	0.11	0.55	**0.62**
Memory used	–	64 B	128 B	256 B	512 B	1024 B	2 B	3 B

Best values of each line are highlighted and are presented in bold face.

syllables. However, it found many more segments than the actual number, which decreases the precision increasing the error rate. By using the mode filter, EZ-IMF was able to find the correct number of segments (reducing AEER except for the species *Rhinella G.*), although it was not able to detect the beginning of the segments as precisely as EZ-I (which resulted in slightly worse figures for P and R).

The TPR and FNR are good idicators of signal loss (considering a point-to-point evaluation). In these results, the values of FNR are consistent with the R (recall), indicating more signal loss for the windowing approaches. The low recall in the case of $N = 1024$ indicates that this value is much larger than the length of syllables, which may result in the loss of the whole syllable in some cases.

Table 3 presents the results achieved by EZ-IMF and EZ-WIN[3] for each species. It is clear that the most challenging species were *Osteocephalus O.* and *Rhinella G.* By carefully inspecting the segmentations we note that even when both methods find more syllables (false positives) than the perfect segmentation (large AEER), EZ-IMF tends to find them within actual syllables which leads to smaller F1 values.

Now let us evaluate a complete system to classify sound, showing the impact of segmentation on the species recognition.

8.1. Impacts of segmentation

We compute how the proportion of correct segmentations affects the final accuracy by Eqs. (14)–(17). Table 4 shows the final results of three metrics. For each segmentation technique, we obtained the final value as the macro-average of all species. The last line is the amount of memory used for compute the segmentation.

In this table, PS (perfect segmentation) column represents the performance of kNN to classify segments defined by a human expert, i.e., with neither FP_s nor FN_s. The last row of Table 4 shows that when we increase the amount of memory, results are not necessarily better. Best values are presented in bold face. As we can see, our segmentation technique outperformed the baselines, considering the F1 metric, which leverages precision and recall. For the baselines, the best result was obtained with a window of size $N = 256$.

9. Conclusions and final comments

The approach presented in this paper is based on the idea of incremental processing with a minimal usage of resources. We demonstrate that our method is effective in detecting audio segments of interest. By segmenting the signal before transmission, the method reduces the amount of data sent by the sensors to the sink node. As transmission is the most energy intensive task, this reduction leads to longer lifetime. Additionally, incremental calculation is suitable for a big data context, in which sounds are received by sink nodes as data streams to be processed in an online fashion (Gama & Gaber, 2007).

When we use sliding windows, it is difficult determine the best size. This size may be chosen by using the evaluation methodology we presented in this work. A small window size reacts quickly to changes being more sensitive to noise, which leads to more false positives. A larger window size loses more syllables, increasing the false negative rate by failing to quickly adapt to changes. As our method is able to continuously adapt to changes, through means of a forgetting process, it performs better in our application.

Because segmentation techniques can retrieve very different amounts of syllables, using classification accuracy for overall evaluation can be misleading. Thus, we proposed a more robust evaluation methodology to assess the performance of a whole system composed by a segmenter and a classifier. This solution allowed us to better quantify the impact of the segmenter on the final system performance.

Our main contributions are threefold: (1) improvements in recall and F-score of automatic segmentation, with few syllable losses; (2) lower dependency on memory (O(1)) and real time response; and (3) a multilevel classifier evaluation method for the entire bioacustical framework. In addition, the proposed method is able to better adapt to gradual changes and yields fewer false negatives taking advantage of the ability to evolve over time.

As future work, we intend to study how to automatically find, in real time, optimal values for α and threshold parameters T_E e T_Z, without compromising the incremental characteristics of our method. Also, we intend to develop and expand our evaluation method to cover more general cases, with more than two classification levels. Our work clearly take advantage of simple characteristics to distinguish noise from signal in an unsupervised fashion. Thus, we will verify how to derive incremental versions of other robust features used in literature.

[3] We chose $N = 512$ for Table 3, as it performed well in Table 2, considering the compromise between F1 and AEER.

Acknowledgements

The authors acknowledge the support granted by FAPEAM through process number 2210.UNI175.3532.03022011 (Anura Project – FAPEAM/CNPq PRONEX 023/2009).

References

Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). Wireless sensor networks: A survey. *Computer Networks*, 38.

Carey, C., Heyer, W. R., Wilkinson, J., Alford, R. A., Arntzen, J. W., Halliday, T., et al. (2001). Amphibian declines and environmental change: Use of remote-sensing data to identify environmental correlates. *Conservation Biology*, 15.

Cheng, J., Sun, Y., & Ji, L. (2010). A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. *Pattern Recognition, 43*, 3846–3852.

Colonna, J. G., Cristo, M. A. P., & Nakamura, E. F. (2014). A distribute approach for classifying anuran species based on their calls. In *22nd international conference on pattern recognition*.

Colonna, J. G., Ribas, A. D., dos Santos, E. M., & Nakamure, E. F. (2012). Feature subset selection for automatically classifying anuran calls using sensor networks. In *International joint conference on neural networks (IJCNN)* (pp. 1–8). IEEE.

Evangelista, T. L. F., Priolli, T. M., Silla, C. N., Angelico, B. A., & Kaestner, C. A. A. . In *Automatic segmentation of audio signals for bird species identification* (pp. 223–228). IEEE.

Fagerlund, S. (2007). Bird species recognition using support vector machines. *EURASIP Journal of Applied Signal Processing, 2007*. 64–64.

Finch, T. (2009). *Incremental calculation of weighted mean and variance. Technical Report*. University of Cambridge Computing Service.

Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. *IEEE International Conference on Multimedia and Expo (ICME)* (Vol. 1, pp. 452–455). IEEE.

Gama, J., & Gaber, M. M. (2007). *Learning from data streams*. Springer.

Garcia, N., Marcias-Toro, E., Vargas-Bonilla, J. F., Daza, J. M., & López, J. D. (2014). Segmentation of bio-signals in field recordings using fundamental frequency detection. In *3rd International work conference on bioinspired intelligence (IWOBI)*. IEEE.

Giannakopoulos, T., Pikrakis, A., & Theodoridis, S. (2008). A novel efficient approach for audio segmentation. In *19th International conference on pattern recognition, (ICPR)* (pp. 1–4).

Giannoulis, D., Stowell, D., Benetos, E., Rossignol, M., Lagrange, M., & Plumbley, M. D. (2013). A database and challenge for acoustic scene classification and event detection. In *European signal processing conference*.

Han, N. C., Muniandy, S. V., & Dayou, J. (2011). Acoustic classification of australian anurans based on hybrid spectral-entropy approach. *Applied Acoustics, 72*, 639–645.

Härmä, A. (2003). Automatic identification of bird species based on sinusoidal modeling of syllables. *IEEE international conference on acoustics, speech, and signal processing (ICASSP'03)* (Vol. 5). IEEE (pp. V–545).

Hu, W., Bulusu, N., Chou, C. T., Jha, S., Taylor, A., & Tran, V. N. (2009). Design and evaluation of a hybrid sensor network for cane toad monitoring. *ACM Transactions Sensors Network, 5*, 4:1–4:28.

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

Jaafar, H., Ramli, D., & Shahrudin, S. (2013). Mfcc based frog identification system in noisy environment. In *International conference on signal and image processing applications (ICSIPA)* (pp. 123–127). IEEE.

Jaafar, H., & Ramli, D. A. (2013). Automatic syllables segmentation for frog identification system. In *9th International colloquium on signal processing and its applications (CSPA)* (pp. 224–228). IEEE.

Jaber, G. (2013). *An approach for online learning in the presence of concept change* (Ph.D. thesis). LIMSI-CNRS.

Khan, S., Pathan, A. K., & Alrajeh, N. A. (2012). *Wireless sensor networks: Current status and future trends*. CRC Press.

Nakamura, E. F., Loureiro, A. A. F., Boukerche, A., & Zomaya, A. Y. (2014). Localized algorithms for information fusion in resource constrained networks. *Information Fusion, 15*, 2–4.

Nakamura, E. F., Loureiro, A. A. F., & Frery, A. C. (2007). Information fusion for wireless sensor networks: Methods, models, and classifications. *ACM Computing Surveys*, 39.

Neal, L., Briggs, F., Raich, R., & Fern, X. Z. (2011). Time-frequency segmentation of bird song in noisy acoustic environments. In *International conference on acoustics, speech and signal processing (ICASSP)* (pp. 2012–2015). IEEE.

Potamitis, I., Ntalampiras, S., Jahn, O., & Riede, K. (2014). Automatic bird sound detection in long real-field recordings: Applications and tools. *Applied Acoustics, 80*, 1–9.

Powers, D. M. W. (2007). *Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation. Technical Report*. SIE-07-001 School of Informatics and Engineering, Flinders University.

Rahman, M., & Bhuiyan, A. (2012). Continuous bangla speech segmentation using shortterm speech features extraction approaches. *International Journal of Advanced Computer Sciences and Applications, 3*, 11.

Ribas, A. D., Colonna, J. G., Figueiredo, C. M. S., & Nakamura, E. F. (2012). Similarity clustering for data fusion in wireless sensor networks using k-means. In *International joint conference on neural networks (IJCNN)* (pp. 1–7).

Rickwood, P., & Taylor, A. (2008). Methods for automatically analyzing humpback song units. *The Journal of the Acoustical Society of America, 123*, 1763–1772.

Sarkar, A., & Sreenivas, T. V. (2005). Automatic speech segmentation using average level crossing rate information. *International conference on acoustics, speech, and signal processing (ICASSP)* (Vol. 1, pp. 397–400). IEEE.

Scaiano, M., & Inkpen, D. (2012). Getting more from segmentation evaluation. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*. NAACL HLT '12 (pp. 362–366).

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*, 427–437.

Theodorou, T., Mporas, I., & Fakotakis, N. (2014). An overview of automatic audio segmentation. *International Journal of Information Technology and Computer Science (IJITCS), 6*, 1–9.

Wang, H., Elson, J., Girod, L., Estrin, D., Yao, K., & Vanderberge, L. (2003). Target classification and localization in habitat monitoring. In *Proceedings of the international conference on speech and signal processing*. IEEE.

Xie, J., Towsey, M., Truskinger, A., Eichinski, P., Zhang, J., & Roe, P. (2015). Acoustic classification of australian anurans using syllable features. In *Tenth international conference on intelligent sensors, sensor networks and information processing (IEEE ISSNIP 2015)*.