Expert Systems With Applications 106 (2018) 107–120

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Feature evaluation for unsupervised bioacoustic signal segmentation of anuran calls

Juan G. Colonna[a,*], Eduardo F. Nakamura[a,b], Osvaldo A. Rosso[c,d,e]

[a] Instituto de Computação (Icomp), Universidade Federal do Amazonas (UFAM), Av. General Rodrigo Octávio 6200, Manaus, AM, CEP: 69077-000, Brasil
[b] Department of Computer Science and Engineering, Texas A&M University, USA
[c] Instituto de Física, Universidade Federal de Alagoas (UFAL), Av. Lourival Melo Mota, S/N - Tabuleiro do Martins, Maceió, AL, CEP: 57072-970, Brasil
[d] Departamento de Informática en Salud, Hospital Italiano de Buenos Aires and CONICET, C.A.B.A., C1199ABB, Argentina
[e] Complex Systems Group, Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Santiago 12455, Chile

ARTICLE INFO

Article history:
Received 11 August 2017
Revised 27 March 2018
Accepted 29 March 2018
Available online 5 April 2018

Keywords:
Unsupervised bioacoustics signal segmentation
Information theory
Permutation entropy
Colored noise

ABSTRACT

We present a comprehensive study of temporal Low-Level acoustic Descriptors (LLDs) to automatically segment anuran calls in audio streams. The acoustic segmentation, or syllable extraction, is a key task shared by most of the bioacoustical species recognition systems. Consequently, the syllable extraction has a direct impact on the classification rate. In this work, we assess several new entropy measures including the recently developed Permutation Entropy, Weighted Permutation Entropy, and Permutation Min-Entropy, and compare them to the classical Energy, Zero Crossing Rate and Spectral Entropy. In addition, we propose an algorithm to estimate the optimal segmentation threshold value used to separate deterministic segments from stochastic ones avoiding the creation of thin clusters. To assess the performance of our segmentation approach, we applied a frame-by-frame, a point-to-point and an event-to-event comparisons. We show that in a scenario with severe noise conditions (SNR $\leq$ 0dB), simple entropy descriptors are robust, achieving 97% of segmentation performance, while keeping a low computational cost. We conclude that there is no LLD that is suitable for all scenarios, and we must adopt multiple or different LLDs, depending on the expected noise conditions.

1. Introduction

The loss of amphibian biodiversity is a worldwide concern. Anuran (frogs and toads) have a close relationship with the environment. By monitoring anuran populations, we can detect ecological stress in early stages (Carey et al., 2001; Cole, Bustamante, Reinoso, & Funk, 2014; Luque, Romero-Lemos, Carrasco, & Barbancho, 2017). The variations in anuran populations can help us understand what is happening in their environment. Most of the monitoring programs are based on acoustic surveys applied by a group of experts and collaborators, who move from one place to another while counting the species and individuals (Gibbs, Whiteleather, & Schueler, 2005; MacKenzie, Nichols, Hines, Knutson, & Franklin, 2003). The full study takes many years and demands a lot of human and economic resources.

One possible solution to mitigate that cost is the development of an automatic method to detect the presence of different anuran species through their calls, without human intervention. In this context, the problem can be addressed by using Wireless Acoustic Sensor Networks (WASNs) (Colonna, Cristo, & Nakamura, 2014; Colonna, Ribas, Santos, & Nakamura, 2012; Ribas, Colonna, Figueiredo, & Nakamura, 2012) and Machine Learning classification techniques to detect the presence of particular species (Brandes, 2008; Colonna et al., 2016; Somervuo, Harma, & Fagerlund, 2006). However, the low cost of this technology results in hardware and software resource constraints, which demand algorithmic solutions of lower computational cost (Nakamura, Loureiro, Boukerche, & Zomaya, 2014; Nakamura, Loureiro, & Frery, 2007).

In the context of WASNs, the sound acquisition is performed non-intrusively by the sensor nodes, which allow us to monitor the environment for a long-term period. Replacing the sensor batteries may be too expensive or even unfeasible. Hence, we need to develop efficient methods that minimize the amount of information being processed, transmitted, or recorded, by the sensor nodes.

To enable monitoring with WASNs, it is necessary to embed an Automatic Call Recognition (ACR) method into the sensor nodes. A general ACR method for recognizing frog species, based on their calls, is shown in Fig. 1. This method consists of three major

* Corresponding author.
E-mail addresses: juancolonna@icomp.ufam.edu.br (J.G. Colonna), nakamura@tamu.edu (E.F. Nakamura), oarosso@gmail.com (O.A. Rosso).

https://doi.org/10.1016/j.eswa.2018.03.062
0957-4174/

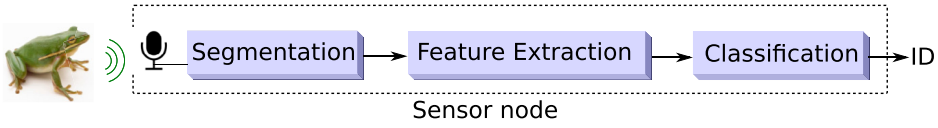


Fig. 1. Automatic Call Recognition Framework (ACR).

processing blocks. The first block performs the acoustic signal segmentation, recognizing the start and end of a minor vocalization unit, named *syllable* (Huang, Yang, Yang, & Chen, 2009; Somervuo et al., 2006). The second block maps the syllable into a feature vector (Brandes, 2008). The last block is a pattern-matching algorithm that considers the input feature vector and a feature set representing all the species included in the reference dataset (Colonna et al., 2014; McIlraith & Card, 1997; Wichern, Xue, Thornburg, Mechtley, & Spanias, 2010). Note that the classification step is not covered in this article.

As we can observe, the segmentation block impacts on the final species recognition rate, i.e., the better the segmentation result, greater is the probability of the correct species recognition (Alonso et al., 2017; Colonna, Cristo, Salvatierra, & Nakamura, 2015; Jaafar, Ramli, & Shahrudin, 2013). Moreover, given the limitations of the sensors, keeping the segmentation method as economical as possible, from the viewpoint of computational complexity, is our major challenge. Therefore, here we provide a solid acoustic descriptor evaluation and comparison for bioacoustic signal segmentation that detects and extracts the syllables of an anuran call in an unsupervised manner.

In real situations, such as rain forests, the scenarios can be complex and present a high acoustic richness, as a result of the interaction of several species at the same place (Depraetere et al., 2012). Therefore, it is not possible to know all possible signal patterns *a priori*. Hence, we propose a change in the segmentation paradigm: instead of trying to identify different signal patterns, we identify only noise segments. Thus, the remaining segments may be considered syllables (see Section 2). This is possible because in our formulation the segmentation task is equivalent to an unsupervised binary classifier, in which we separate features that belong to segments of either “signal” class or “noise” class. After the segmentation, the final classifier (third block of Fig. 1) is responsible for the species recognition. The impact of the segmentation on the final recognition rate has been studied (Colonna et al., 2015; Jaafar et al., 2013; Somervuo et al., 2006), but nothing was reported about individual LLDs applied to segmentation. The classification step is out of scope of this work.

We present an unsupervised segmentation approach. We focus our experiments on using only a reduced set of Low-Level acoustic Descriptors (LLDs) from temporal and spectral domains to cope with the hardware restrictions of low-cost sensors. Moreover, our method is useful for segmenting calls stored into a bioacoustic database in an unsupervised manner. We then analyze the segmentation performance considering several noise conditions, including white and colored noises (blue, red, violet and pink). In literature few authors discuss the problem of such color noises, but given the goal of our application it is essential.

The contributions of this work are twofold:

1. a comparative assessment of three unconventional LLDs based on the new Permutation Entropy (PE) methodology and its variants (Weighted Permutation Entropy - WPE and Permutation Min-Entropy - PME), one LLD based on Spectral Entropy (H_{FFT}), and two common temporal LLDs (Energy - E and Zero Crossing Rate - ZCR); and
2. an algorithm to find the optimal segmentation threshold for the syllables using the descriptors mentioned above.

Hence, we perform several evaluations trying to answer why the same features, such as E, ZCR and H_{FFT}, are often used in the literature, even in situations where the noises may have different spectral characteristics. To evaluate the performance of our algorithm to find the best threshold, we compared it against the Otsu and k-Means methods. To the best of our knowledge, this is the first work that applies and compares the new Permutation Entropy methodology, and its variants, to segment bioacoustic signals.

Finally, the performance was quantified by computing: the Area Under the ROC Curve (AUC), the Acoustic Event Error Rate (AEER), the false positive and false negative rates (FPR and FNR), the F-Score (F1), and the accuracy (Acc). Then, supported by experimental results, we demonstrate that these entropy quantifiers are robust enough for real applications, even considering noise levels below than 0 dB, achieving an accuracy superior to 95%. These results are significant to those who intend to design and implement a non-intrusive environmental monitoring method.

All these evaluations are equally important to obtain the final performance of the segmentation approach. Each metric helps to highlight different aspects of the segmentation. A complete assessment is generally not considered in the related works, in which the quality of the segmentation is frequently evaluated through the classification rate of the species. The problem with this is that the classifier also produces errors that mask the segmentation errors. This makes it difficult to identify the real failures of the whole system.

The remainder of this paper is organized as follows. Section 2 defines the segmentation problem of bioacoustic calls. Related works are presented in Section 3. Section 4 describes the set of LLD assessed in our comparative study. The algorithm we propose to find the optimal threshold value is presented in Section 5. We also show three different performance assessments of the segmentation task in the Sections 6.1, 6.2 and 6.3. Additionally, given that the LLDs are not necessarily correlated, in Section 6.4 we show a ranking of LLDs based on Information Gain criterion and an evaluation of some LLD combinations. Section 7 discusses which are the most robust LLDs to segment anuran calls.

2. Problem description

The accuracy of species recognition depends on two major factors: (1) the classifier's ability to separate different signal patterns represented in the feature space; and (2) the quality of the mapping function, which transforms the raw signal segments into discriminating features. The quality of features depends on the mapping function, but also depends on using the correct part of the input signal, which contains more useful information. Thus, the final accuracy of the complete system, presented in Fig. 1, depends on the capacity to select the appropriate signal segments from the input.

Although different frog species may have different types of calls for different purposes, such as territory delimitation or mating, in all cases there are small signal patterns repeated along time. These units, known as syllables, are the smallest bioacoustics pattern useful to identify different species. Thus, each species has its

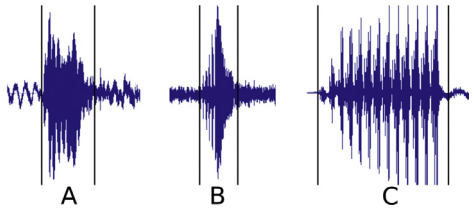


Fig. 2. Three different syllable patterns belonging to the species (a) Adenomera h., (b) Hypsiboas cinerascens and (c) Scinax ruber. The vertical dot lines represent the beginning and end of each segment.

own set of syllables. Fig. 2 depicts three syllables from three different species.

Fig. 3 is a vocalization sample with three consecutive syllables of the species Adenomera h.. Between each syllable, we have the background noise. The segmentation problem consists of recognizing the beginning and the end of each segment that contains only noise, doing this accurately and automatically (background segments on Fig. 3). Consequently, the remaining segments correspond to the syllables that we want to classify. In this approach, we do not need to know all possible syllable patterns *a priori* to be able to extract such syllables.

Formally, the input bioacoustic signal $x(t) = \{x_1, x_2, \ldots, x_N\}$ is a time series in which the values represent the acoustic pressure levels (or amplitude) within $1 \leq t \leq N$, in which N is the maximum signal length. A frame $\mathbf{x}_k = \{x_i, x_{i+1}, \ldots, x_{i+n}\}$ is a subset of size n with consecutively signal values. Thus, a representation of the signal by a frameset can be obtained by a sliding window of size n. The main challenge is how to classify these frames into: "signal" or "noise" (1 or 0). To address this problem, we can represent the frame values $\mathbf{x}_k$ by a set of LLDs. For example, if we use the entropy value of the first frame $\mathbf{x}_1$, we can applied the binary decision rule:

$$\text{class}(\mathbf{x}_1) = \begin{cases} 1 & \text{if } \mathrm{H}(\mathbf{x}_1) \leq T_H \\ 0 & \text{if } \mathrm{H}(\mathbf{x}_1) > T_H \end{cases}, \tag{1}$$

in which T_H is a threshold for the entropy value of each frame. With this rule, we assign the class "signal" to frames of low entropy. Since entropy can be interpreted as a measure of "impurity", the higher the value, the greater the probability that the underlying signal is a random noise. A similar rule can be built for other LLDs.

A secondary challenge, associated with the application of this rule, is how to find the optimal threshold value T_H. The optimal T_H is a trade-off between the sensitivity to the noise and the precision of the syllable boundaries. To accomplish this, we present a binarization technique described in Algorithm 2. The results and experiments of Section 6.1, support the hypothesis that this algorithm produces an optimal frame division.

3. Related work

In bioacoustic monitoring approaches, the recognition task has been discussed and studied extensively. However, the audio segmentation is usually neglected, treated as a secondary task or performed manually (McIlraith & Card, 1997; Strout et al., 2017). For instance, Luque et al. (2017) explain that syllable extraction is a highly complex task, especially in the case of noisy recordings, and they proposed an alternative method based on the processing of successive frames to avoid it. Tomasini, Smart, Menezes, Bush, and Ribeiro (2017) proposed a new set of acoustic features to avoid segmentation. However, segmentation is still very useful for reducing the amount of data transmitted, processed, or stored.

Most of the related works perform a syllable segmentation, but only a few of them are concerned about the Low-level Acoustic Descriptors (or features) employed and the combination of such features to segment anuran calls. The impact of the segmentation is also neglected in frog recognition problems (Colonna et al., 2015; Evangelista, Priolli, Silla, Angelico, & Kaestner, 2014; Jaafar et al., 2013; Lopes, Koerich, Silla, & Kaestner, 2011; Somervuo et al., 2006; Wichern et al., 2010). Lopes et al. (2011) highlighted the importance of segmentation compared to the use of the entire audio in bird species recognition tasks. Evangelista et al. (2014) compared manual segmentation against automatic techniques for bird calls showing gains between 7% and 23% of recognition rate. Jaafar et al. (2013) and Colonna et al. (2015) confirmed that the automatic segmentation can separate the most relevant audio fragments in anuran calls, hence, becoming a fundamental part of the

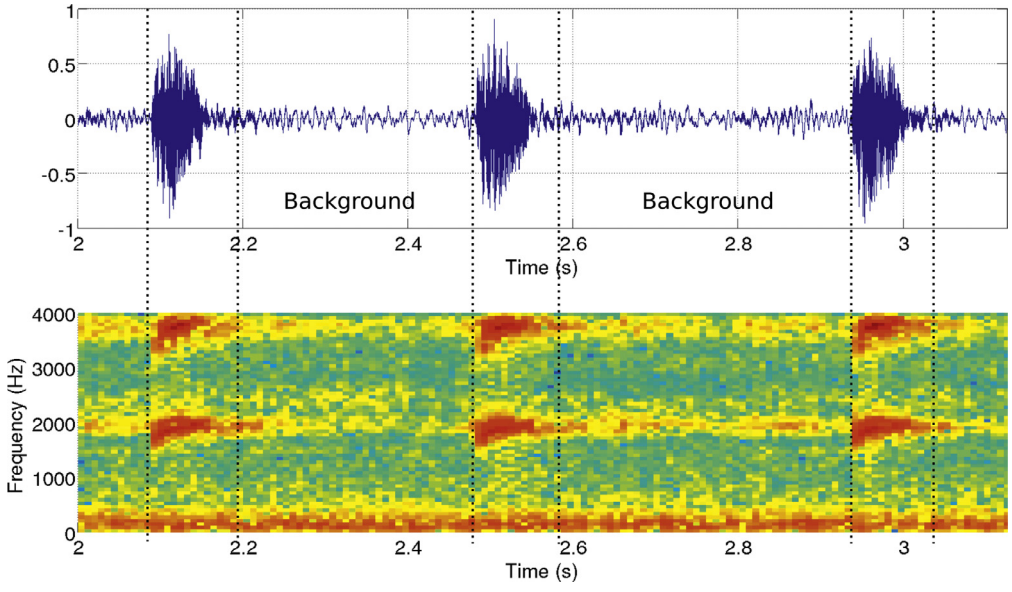


Fig. 3. Three syllables of an Adenomera h. call: signal amplitude (above) and spectrogram (below). The vertical dot lines represent the beginning and end of syllable patterns in time and frequency domains. Note that between the 0 and 300 kHz frequency bands there is background noise present all the time, inside and outside the syllables.

monitoring framework. However, there is no consensus on how we should evaluate the gains over the recognition rate.

We identify that most of the bioacoustic related works use heterogeneous sets of features without performing an appropriate evaluation of each individual feature or even using different feature combinations. Using an extensive set of features becomes unfeasible when the processing (or transmission) capacity is limited by memory or power-saving requirements.

Automatic sound segmentation has been widely studied in other contexts, usually focusing on music and human voice streams (Foote, 2000; Giannakopoulos, Pikrakis, & Theodoridis, 2008; Sarkar & Sreenivas, 2005; Theodorou, Mporas, & Fakotakis, 2014). In such studies, it is common to prioritize robust solutions even when they lead to higher costs, being impractical for a low-cost sensor. Expensive approaches were also employed in the study of bioacoustic signals as digital image processing from its spectrogram (Bardeli, 2009). For instance, morphological filters can be applied over the spectrogram by using it as an image, and finding regions of interest comprising neighboring pixels (Aide et al., 2013; Oliveira et al., 2015; Potamitis, 2014; Xie et al., 2015). However, image processing techniques are too expensive for memory to low-cost sensors.

Cettolo, Vescovi, and Rizzi (2005) state that the segmentation models can be classified into three categories:

- ***Energy models:*** in these models the energy of each frame is compared to a threshold. The frames of which energy is higher than the threshold are considered signal, while the others are considered noise (Alonso et al., 2017). The changing point is identified when two successive frames belong to different classes. The signal energy can be obtained in the temporal or spectral domains (Noda, Travieso, & Sánchez-Rodríguez, 2016), and other features such as the Zero Crossing Rate may also be used to detect a change point (Colonna et al., 2015). The energy-based segmentation algorithms can be easily implemented and applied in an unsupervised manner. As the energy of the signal is closely related to the amplitude (or sound pressure level), these models are susceptible to the noise floor level and the impulsive noise (Colonna et al., 2012). A major challenge in these approaches is the choice of the optimal threshold.
- ***Probabilistic models:*** in these models each acoustic class (e.g., music, speech or noise) is represented by the underlying probability distribution function (PDF) of the signal values or their feature values, from which a comparison criterion between PDFs is applied to find the segmentation boundaries. Some common criteria are the *Bayesian Information Criterion* (BIC) (Cettolo et al., 2005; Cheng & Wang, 2003; Heinicke et al., 2015) and the Dynamic Bayesian Network (DBN) (Wichern et al., 2010). Other possibilities include the comparison of segmented regions of different entropy values (Shen, Hung, & Lee, 1998; Wu & Wang, 2005) or the computation of similarity as the *Kullback–Leibler* (KL) divergence or the Maximum Mean Discrepancy (Fagerlund & Laine, 2014; Sinn, Keller, & Chen, 2013). Although these models are more robust than a simple energy detector, finding a suitable PDF is a challenge that usually demands different methodologies.
- ***Explicit models:*** with these models each silence (or signal pattern) in the audio stream is detected by a supervised explicit model that was trained with past examples (Neal, Briggs, Raich, & Fern, 2011). These are the most robust approaches, but the past samples must be labeled by a human expert, consuming more time and human effort. An advantage of these approaches is the possibility of performing segmentation and recognition of species in a single step (Chu & Blumstein, 2011; Ren et al., 2009). This category includes the approaches that employ digital image processing, based on morphological filters, applied to the spectrogram (Aide et al., 2013; Potamitis, 2014). Despite the good results, the usage of morphological filters is computationally expensive.

A similar categorization for the segmentation approaches was given by Theodorou et al. (2014). They used two categories: the *distance-based* and the *model-based* segmentation. The model-based segmentation is similar to the explicit models defined above, these methods use supervised machine learning algorithms. The distance-based segmentation can use Euclidean distance, BIC, KL, Generalized Likelihood Ratio (GLR), and the Hotelling T2 statistic. To improve the segmentation result, the distance-based model allows us to estimate the distance: frame-to-frame, frame-to-group, or group-to-group of frames. These methods are less sensitive to local anomalies of the signal.

We propose Algorithm 2 to find the optimal segmentation threshold and avoid the issues of the energy models. This algorithm employs all the signal frames to find an equidistant value between the averages of the two groups (“signal” and “noise”), maximizing the intra-class distance, while trying to keep the PDFs of the groups balanced. This is a group-to-group approach. In addition, to obtain the signal features, we first apply different PDF methodologies, then we compute the entropy values. Thus, we can consider our method as a hybrid approach that combines the simplicity of the energy models, the discrimination power of the probabilistic models and the robustness of the group-to-group approaches.

4. Fundamentals concepts

This section presents the fundamental knowledge that supports this work.

4.1. Energy and zero crossing rate

The signal’s Energy (E) allows us to know when the signal amplitude increases, while the Zero Crossing Rate (ZCR) provides an approximation of the dominant frequency. These two temporal features, commonly used in bioacoustics processing methods (Jaafar & Ramli, 2013; Jaafar et al., 2013; Rahman & Bhuiyan, 2012) are given by:

$$\mathrm{E} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \text{ and} \tag{2}$$

$$\mathrm{ZCR} = \frac{1}{2n}\sum_{i=1}^{n} |\mathrm{sign}(x_i) - \mathrm{sign}(x_{i-1})|, \tag{3}$$

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

$$\mathrm{sign}(x_i) = \begin{cases} +1, & \text{if } x_i \geq 0; \\ -1, & \text{if } x_i < 0. \end{cases} \tag{4}$$

4.2. Spectral entropy

To compute the spectral entropy, a frame $\mathbf{x}_k$ is first transformed into the spectrum $\mathrm{S}(f) = \mathcal{F}(\mathbf{x}_k)$ by using a Fast Fourier Transform (FFT) (Sueur, Pavoine, Hamerlynck, & Duvail, 2008; Wu & Wang, 2005). This spectrum $\mathrm{S}(f)$ is normalized to obtain a probability mass function:

$$S(f) = \frac{|S(f)|}{\sum_{f=0}^{f_s/2} |S(f)|}, \quad (5)$$

which is used to compute the normalized Spectral Entropy

$$H_{FFT} = -\sum_{f=0}^{f_s/2} \frac{S(f)\ \log(S(f))}{\log(n)}, \quad (6)$$

where n is the length of $S(f)$ and f_s the sampling frequency. In this case, only the positive part of the spectrum is used.

4.3. Permutation entropy

The Permutation Entropy (PE) characterizes the dynamics of a time series (Bandt & Pompe, 2002; Soriano, Zunino, Rosso, Fischer, & Mirasso, 2011). This quantifier, together with the permutation statistical complexity, allows us to compare or distinguish deterministic, stochastic, and chaotic behaviors in time series (Labate, Foresta, Morabito, Palamara, & Morabito, 2013; Rosso, Larrondo, Martin, Plastino, & Fuentes, 2007).

Bandt and Pompe (2002) transform a temporal series of real values into a symbolic representation known as "ordinal patterns" (π_j). Thus, the complete series is represented by a set of symbols $\Pi = \{\pi_1, \pi_2, \ldots, \pi_{m!}\}$, in which m is the length of each pattern (*embedding dimension*). The set Π is constructed with all possible permutations of integers numbers between 1 and m, e.g., for $m = 3$ the symbol set is $\Pi = \{(1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1)\}$.

The numbers in each π_j pattern represent the sequential index of the original real values after being sorted. For example, given a set of three real numbers (0.4, 2.3, 1.5), with indices (1,2,3), after these being sorted in descending order (0.4,1.5,2.3), the resulting index permutation is (1,3,2), which matches the pattern π_2 of set Π. We can summarize the procedure to obtain the histogram of Π as follows:

1. Select m consecutive values of the signal $(x_i, \ldots, x_{i+m})$;
2. Apply $sort(x_i, \ldots, x_{i+m}$, 'descending'), recover the indices of the sorted values, and find the corresponding π_j into Π;
3. Increase the relative frequency of the corresponding pattern and the time index, $f_{\pi_j} = f_{\pi_j} + 1$ and $i = i + 1$;
4. Repeat the steps above until the end of the signal is reached.

The normalized histogram of Π is the frequent approach to obtain the probability associated with each ordinal pattern π_j, and we denoted it by p_{π_j}, from which the normalized permutation entropy can be computed as:

$$PE = -\sum_{j=1}^{m!} \frac{p_{\pi_j} \log(p_{\pi_j})}{\log(m!)} \quad (7)$$

The entire procedure is summarized in Algorithm 1[1]. Here, τ (*time lag*) is an extra parameter that controls the time scale of PE. For example, suppose we analyze the time series $x = \{0.4, 2.3, 1.5, 1.7, 0.5, 1.0\}$ using $m = 3$ and $\tau = 1$, the corresponding pattern are: $sort(0.4, 2.3, 1.5) = (1,3,2) \rightarrow \pi_2$, $sort(2.3, 1.5, 1.7) = (2,3,1) \rightarrow \pi_4$, $sort(1.5, 1.7, 0.5) = (3,1,2) \rightarrow \pi_5$, and $sort(1.7, 0.5, 1.0) = (2,3,1) \rightarrow \pi_4$. If we keep $m = 3$ and change $\tau = 2$, the corresponding pattern will be: $sort(0.4, 1.5, 0.5) = (1,3,2) \rightarrow \pi_2$, and $sort(2.3, 1.7, 1.0) = (3,2,1) \rightarrow \pi_6$. This change permits us to observe the occurrence of π_6, which was not possible with $\tau = 1$. Hence, τ is useful to analyze the behavior of the signal at different time scales (Soriano et al., 2011; Zunino, Soriano, & Rosso, 2012).

[1] A Matlab sample code is available at https://goo.gl/VQvfYm

Algorithm 1 Permutation entropy (PE) calculation.

```
1:  function PE(x, Π, m, tau)
2:      n=length(x);
3:      f=zeros(1,m!);    ▷ Initial frequency of each π_j
4:      for i=1 to n-tau*(m-1) do
5:          [values,indices]=sort(x(i:tau:i+tau*(m-1)))
6:          j = HashTable(indices, Π);    ▷ Pattern index
7:          f(j)=f(j)+1;                  ▷ Increment of π_j
8:      end for
9:      f=f(find(f=0));                    ▷ To avoid NaN
10:     p=f./sum(f);                       ▷ Normalization
11:     return PE=-sum(p.*log(p))*(1/log(m!));
12: end function
```

The last consideration of this methodology is the condition $m! \ll n$ (n is the frame size), which we must satisfy to ensure that we have a high probability of observing all π_j. When the sequence of $\Pi \equiv \{p_j : j = 1, \ldots, m!\}$, generated from the underlying signal, has a uniform histogram, then $PE \approx 1$. In this case, if the condition $m! \ll n$ is satisfied, we conclude that the frame $\mathbf{x}_k$ is a sequence of white noise, i.e., generated by an independent and identically distributed (i.i.d.) random variable. The opposite case, when the histogram of Π is concentrated in one particular π_j, then $PE \approx 0$, and the signal is characterized by a deterministic behavior or has a high trend. Within the interval $0 \leq PE \leq 1$, there are several histograms of Π describing the level of randomness of the signal. In this work, we use this property to separate frames with background noise from syllables.

4.4. The weighted permutation entropy and the permutation min-Entropy

The Weighted Permutation Entropy (WPE) and the Permutation Min-Entropy (PME) are two quantifiers derived from the original PE methodology. The WPE was proposed to put more weight on the patterns that have abrupt amplitude changes (Fadlallah, Chen, Keil, & Príncipe, 2013). To do this, we must replace line 7 in Algorithm 1 by `f(j)=f(j)+var(values)`, where `var(values)` is the variance of the signal values of π. This change arises from the observation that the same ordinal pattern π_j may come from values with a different amplitude, as in our previous example, where the values (2.3,1.5,1.7) and (1.7,0.5,1.0) belong to the same ordinal pattern π_4, but with a different variance.

The PME only considers the π_j among Π with the highest probability (Zunino, Olivares, & Rosso, 2015). To do this, we should replace line 10 in the Algorithm 1 by `p=max(f./sum(f))`, and line 11 by `PME=-log(p)*(1/log(m!))`. This modification holds the original properties of PE and becomes more robust to find deterministic components under high noise conditions.

4.5. Colored noise

Many physical phenomena can produce diverse types of colored noise (ξ) (Lowen & Teich, 1990; Vasseur & Yodzis, 2004). Recordings of acoustic signals are not an exception (Voss & Clarke, 1978). In this cases, the noise time-series may be characterized by a function with Power Spectral Density (PSD) that obeys a power law of the form (Kasdin, 1995):

$$\xi(f) = \frac{L(f)}{|f|^{\alpha}}, \quad (8)$$

in which the exponent α is a real number within $[-2, 2]$, and $L(f)$ is a constant proportional to the process variance

(Kasdin, 1995; Plaszczynski, 2007). The following α values determine some common types of noise:

- $\alpha = 0$ models the white noise containing equal amount of energy in all frequency bands;
- $\alpha = 1$ models the pink noise with equal sound pressure level in each octave band decreasing the energy as the frequency increases;
- $\alpha = 2$ models the red (or brown) noise, which is common in oceanographic recordings, it describes the ambient underwater noise from distant sources (Rudnick & Davis, 2003);
- $\alpha = -1$ models the blue noise that contains more energy as the frequency increases (Ballón et al., 2011); and
- $\alpha = -2$ models the violet noise, which further increases the energy at high frequencies.

The Signal-to-Noise Ratio (SNR) can be obtained by:

$$\mathrm{SNR} = 20\log_{10}\left(\frac{\sigma_x}{\sigma_\xi}\right), \tag{9}$$

in which σ_x and σ_ξ are the standard deviations of the original signal and the added noise, respectively. By varying this ratio, we simulate different distances between the animal, which produces the call, and the sensor, which records the sounds.

4.6. Peak noise

Peak noise is the sparse occurrence of impulses (high energy and short duration). Typically, these impulses are denoted by $\pm\delta(i-k)$, in which i stands for temporal index of x_i and k the temporal position of δ. In a signal with normalized amplitude within the range $[-1, 1]$, $\delta(\bullet)$ may randomly assume the maximum or the minimum values ± 1. The occurrence time k is a uniform random variable, and the noise density is the ratio between the total amount of impulses K, added to the signal, and the length of the signal (N):

$$d_\delta = \frac{1}{N}\sum_{\forall k\in K}|\delta(i-k)|. \tag{10}$$

This is an extremely uncorrelated noise condition and may appear due to several causes, e.g., electrical discharges or loud noises due to weather or environmental conditions.

4.7. The ROC curve

In this work, we use the Receiver Operating Characteristic (ROC) curve to summarize the performance of our binary classifiers. The ROC curve is generated by plotting the True Positive Rate (TPR) with the False Positive Rate (FPR) while varying the decision threshold level between zero and one, and assigning a class to each observation, which is compared to the correct class. Thus, the detection conditional probability becomes a function of the false-alarm probability and help us select the possibly optimal model, independently of the class distribution (Fawcett, 2006; Slaby, 2007).

We also use the Area Under ROC curve (AUC) as an accuracy measurement of a given test. This metric is suitable to understand the ROC plot as a single scalar value which indicates the performance of the classifier. The most important property of AUC is that it represents the probability of randomly selecting a pair of instances (positive and negative).

4.8. Metrics based on decision table

In addition to the ROC analysis, which performs a comparison by frames, we need to know the number of missing syllable's points. By comparing each point of the segmented signal to the true segmentation, and considering this as a binary variable, we can build a decision table. Hence, the result can be summarized by the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), and traditional metrics such as Precision (Pre), Recall (Rec), and F-Score (F1) can be computed. These metrics are defined as:

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{11}$$

$$\mathrm{Rec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{12}$$

and

$$\mathrm{F1} = \frac{2\,\mathrm{Pre}\,\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}}. \tag{13}$$

From the decision table, we also derive the False Negative Rate (FNR), or miss rate; the False Positive Rate (FPR), or false alarm; and the accuracy (Acc):

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}}, \tag{14}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \tag{15}$$

and

$$\mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}}. \tag{16}$$

5. Experimental methodology

The manual signal segmentation of the database is a crucial step to provide the Ground Truth (GT) and to access the methods' performance. We collected the audio of fourteen different frog species with 3155 syllables and 6324 segments, which were manually labeled by a human expert into two classes: "signal" or "background noise"[2] (BN). These recordings were collected *in situ* in the Amazon rainforest, in geographic areas around the Federal University of Amazonas. All of them were recorded without compression in raw format (*.wav*) with a sampling frequency of 44.1 kHz. No signal filter was applied. Table 1 presents the species (first column) and their respective number of syllables (second column).

Every call was mapped to a frame set (S). The frame length chosen was 23.21 ms ($n = 1024$ points) to cope with the PE condition $m! \ll n$, in which $m = 4$ and $\tau = 1$. We decided not to use overlapping to avoid counting repeated points, which would result in an artificial improvement in the segmentation performance. Each frame was represented by a feature value of those described in Section 4, and its timestamp $S(f, t)$. As a result, we obtained a new temporal series within the feature space $S_{id} = \{S(f, t_0), S(f, t_1), \ldots, S(f, t_n)\}$ in which id stands for the ID of the call. Fig. 4a depicts an anuran call (in gray) and its representation through different features.

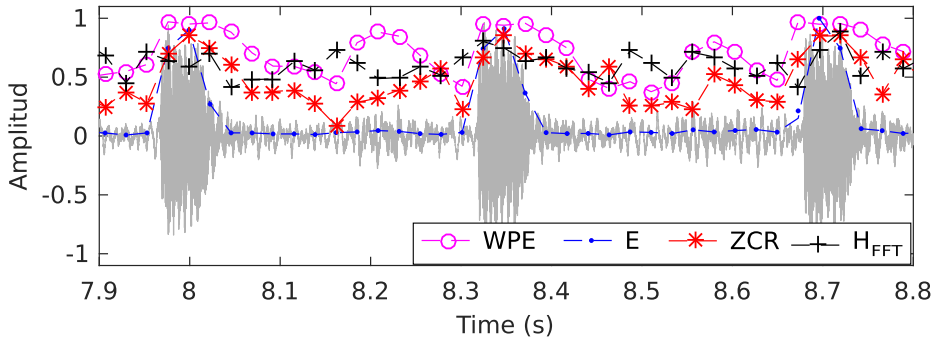
We normalize the values of S into the interval $0 \le \hat{S} \le 1$, by applying the equation:

$$\hat{S}_{id} = \frac{S_{id} - \min(S_{id})}{\max(S_{id}) - \min(S_{id})}, \tag{17}$$

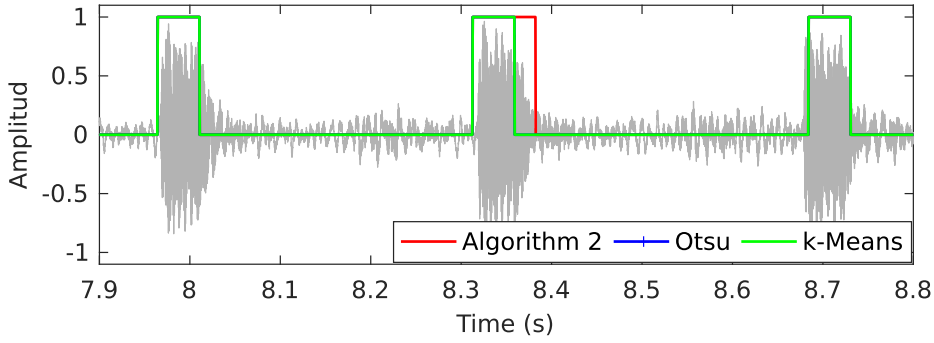
to E and ZCR features, and $\hat{S}_{id} = 1 - \hat{S}_{id}$ to the entropy-based features (PE, WPE, PME and $\mathrm{H_{FFT}}$). This normalization allows us to use the feature values as output scores retrieved by a binary classifier.

The Energy models explained in Section 3 require finding an optimal segmentation threshold. The procedure we propose to find that threshold is presented in Algorithm 2. This algorithm divides the frameset ($\hat{S}_{id}$) into two groups of maximum separation

[2] Our annotated dataset is available at http://bit.ly/1Kd6jYx.



(a) Anuran call represented by a temporal series of features.



(b) Segmentation results using only E.

Fig. 4. A call of the *Adenomera h.* species in gray: (a) feature values of the signal frame-by-frame using: E, WPE, H_{FFT}, and ZCR; (b) an example of segmentation using only E values and applying the threshold computed by the proposed algorithm, the Otsu method and a binary k-Means. Note that the segmentations corresponding to Otsu and k-Means are superimposed.

Algorithm 2 Optimal threshold selection.

```
1: function THRESHOLD(Ŝ)
2:     T_i=0; m_1=0; m_2=0; T_f=mean(Ŝ);
3:     while |T_f − T_i| > 0.01 do
4:         m_1=mean(Ŝ ≤ Tf);
5:         m_2=mean(Ŝ > Tf);
6:         T_i=T_f;
7:         T_f=(m_1+m_2)/2;
8:     end while
9:     return T_f;
10: end function
```

between their means trying to maximize the inter-class distance (lines 4 through 7). Note that Algorithm 2 does not require the frame labels as input, it just needs the entropy values or any other feature representing each frame.

Different from other clustering our method attempts to balance the PDFs of the resulting sets to avoid creating thin clusters with few samples. This procedure is based on an optimal image binarization technique (Sezgin & Sankur, 2004). Once we find the threshold, a simple comparison rule is applied to decide whether it is signal ($\hat{S}_{id}(f,t) \geq T_f$) or background noise ($\hat{S}_{id}(f,t) < T_f$). The relation between false positives and true positives given by this rule is illustrated in Fig. 5a.

Given the time-sparse characteristics of the syllables in anuran calls, there is usually a much larger number of frames corresponding to silences (or background) than frames of syllables. This class imbalance causes traditional threshold selection methods, such as the Otsu method (Yuan, Martínez, Eckert, & López-Santidrián, 2016) or clustering techniques like k-Means (Kamper, Livescu, & Goldwater, 2017), fail to losex a larger number of signal frames.

Fig. 4b shows a segmentation example of an *Adenomera h.* call using only the energy of the signal. This example depicts that the Otsu and k-Means methods produce higher thresholds which cause the partial loss of a syllable. Regarding this problem, we performed several tests and we found that these two

Table 1
Dataset description. The first two columns list the species analyzed and their number of syllables identified by a human expert (Ground Truth (GT)). The other columns present the number of syllables retrieved by each LLD we assessed: Energy (E), Permutation Entropy (PE), Weighted Permutation Entropy (WPE), Permutation Min-Entropy (PME), Zero Crossing Rate (ZCR), and Spectral Entropy (H_{FFT}).

		Features					
Species	GT	E	PE	WPE	PME	ZCR	H_{FFT}
Adenomera h.	58	57	83	91	94	155	158
Hyla m.	39	51	93	89	97	217	12
Adenomera a.	50	50	193	164	194	168	156
Ameerega t.	86	92	105	99	104	128	0
Osteocephalus o.	26	33	310	248	323	324	582
Rhinella g.	2	3	2	3	2	66	0
Scinax r.	57	27	17	34	20	58	0
Hypsiboas c.	1548	1403	2533	1941	2971	4021	2233
Brachycephalus e.	1184	116	132	115	131	2	0
Aplastodiscus albof.	28	28	147	133	151	125	0
Aplastodiscus albos.	8	7	155	181	215	158	3
Aplastodiscus p.	13	13	13	13	24	110	0
Dendropsophus a.	49	46	256	194	213	182	2
Dendropsophus e.	7	7	75	81	123	1	0
Total	3155	1933	4114	3386	4662	5715	3146

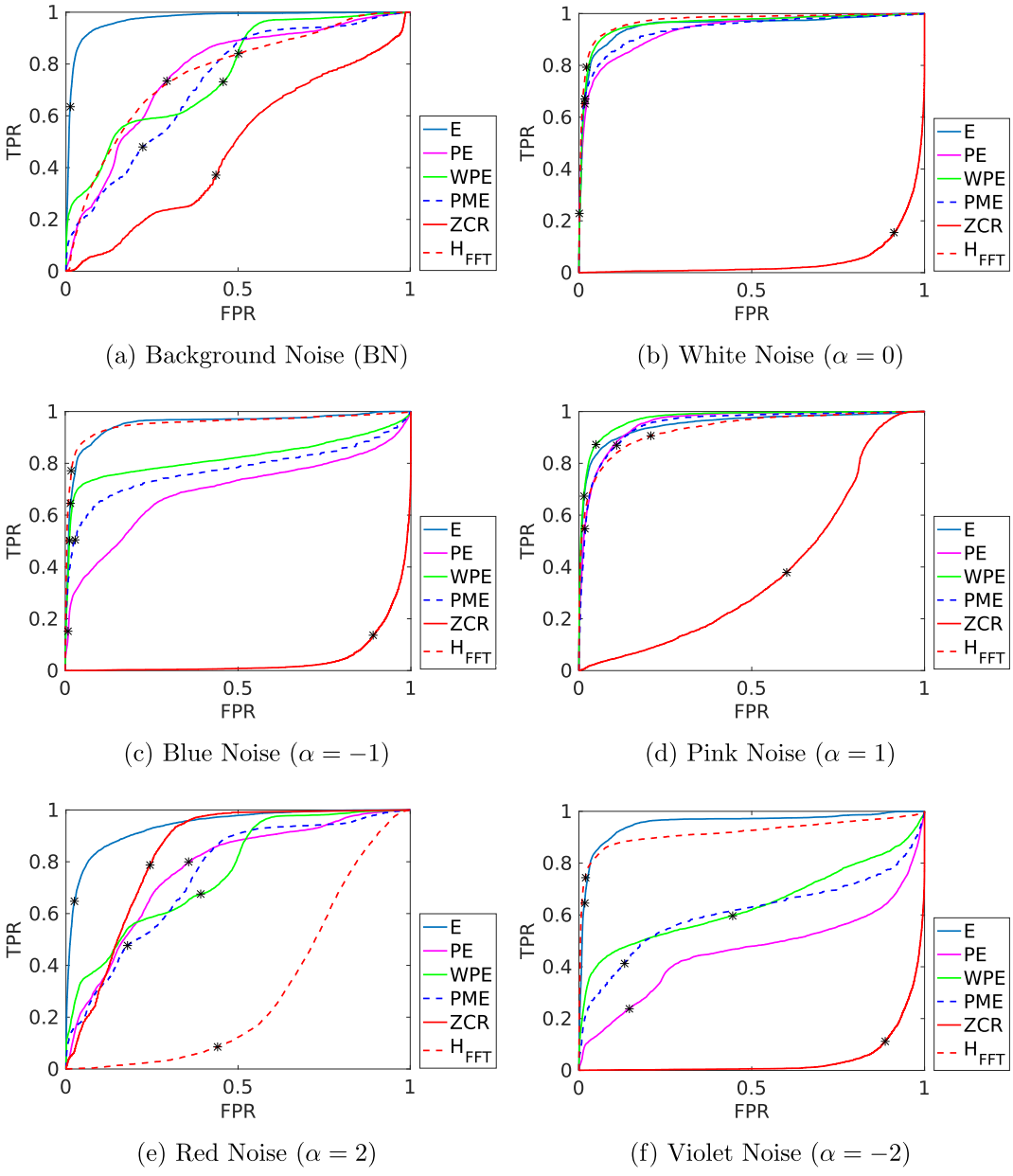


Fig. 5. ROC curves showing the segmentation behavior of each feature adding different noise profiles at 0 dB. The black dots (∗) exemplify the output produced by the rule of Eq. (1) combined with Algorithm 2. The notation for the different LLD used is the same as in Table 1.

methods produce slightly worse results than those obtained with our proposed algorithm. Additionally, we also notice that the random initialization of the k-Mean centroids might produce an instability in the resulting segmentation. For these reasons in the next sessions we will analyze the segmentation results using only Algorithm 2.

In the experiments presented in the following sections, we employ the temporal features E and ZCR, and the spectral feature H_{FFT} as our baselines. The signal energy was used by Alonso et al. (2017) to segment the anuran calls, a combination of E and ZCR was adopted by Colonna et al. (2015), and finally, Wu and Wang (2005) have applied H_{FFT} for endpoint detection in noisy environments. Therefore, comparing these features with the set of features based on Permutation Entropy is equivalent to comparing against the baseline approaches cited.

6. Results

In the previous section we present a segmentation example with the proposed method. In this section we analyze a larger set of species. To determine which are the most appropriate features, we used different metrics to evaluate the quality of segmentation: (a) a frame-by-frame metric quantified by receiver operating characteristic (ROC) curves and the AUC values; (b) an event-to-event metric, called Acoustic Event Error Rate (AEER); and (c) a point-to-point metric that account errors by measuring the Precision, Recall and F-Score (F1).

6.1. Frame-by-frame analysis

To segment syllables, we need to determine whether a frame of $\hat{S}_{id}$ should be transmitted based on its feature values. Thus, we

Table 2
The performance of each feature quantified through the AUC with different noise types. BN stands for background noise. The α parameter is from Eq. (8). Numbers in bold represent the best performance values of each column.

	BN	$\alpha=0$	$\alpha=-1$	$\alpha=1$	$\alpha=2$	$\alpha=-2$
E	**0.97**	0.95	**0.95**	0.95	**0.93**	**0.95**
PE	0.76	0.93	0.69	0.95	0.77	0.46
WPE	0.76	0.96	0.81	**0.97**	0.77	0.62
PME	0.72	0.94	0.77	0.94	0.73	0.61
ZCR	0.47	0.04	0.04	0.38	0.84	0.04
H_{FFT}	0.76	**0.97**	**0.95**	0.93	0.32	0.91

carried out an experiment in which we assign the positive class (+1) to each frame if at least 30% of its points belong to a ground truth syllable. After that, we use the feature value $\hat{S}_{id}$ as the classifier score to plot the ROC curve. With this curve, we calculate the AUC to access the segmentation performance. Fig. 5a shows the ROC curves for all the species listed in the Table 1.

In a real environment, we might face different noise patterns. Hence, we contaminated the original dataset by adding: white, blue, pink, red, and violet noise, with the same variance value as the original signals. Fig. 5 shows the performance of the segmentation by adding different artificial noise at 0 dB, using Eq. (8). The black dots (∗) represent the FPR-TPR relation given by the threshold found with Algorithm 2. Thus, as we expect, the threshold produces low FPR and high TPR for the best curves, i.e., the curves that maximize the AUC. It should be stressed that the construction of the ROC curves takes into account all the FPR-TPR rates produced by all possible thresholds $T_H \in [0, 1]$.

The performance of all features is shown in the BN column of Table 2. These values indicate that, for this scenarios, E has a better performance separating the amplitude of the syllables in the original signals. The remaining columns show the AUC for background noise with the addition of different types of colored noise ($y = x + \xi$), resulting in a Signal-to-Noise Ratio (SNR) of 0 dB. The segmentation based on entropy measures is suitable when we have random noise, being able to better differentiate frames with deterministic patterns from those with stochastic patterns.

The ZCR has the poorest performance due to its sensitivity to frequency changes. Since it is a rough approximation of the main frequency, any spectral perturbation may produce a negative impact on its value. In cases with white, blue, or violet noise (especially those with high spectral energy into upper frequencies), the inversion of the ZCR rule can produce a good result. However, the ZCR is impractical for a real application due the fact that is not possible to know the contamination *a priori* or invert the rule when the acoustic scenario changes.

In Fig. 6, we show the AUC variation as a function of σ_ξ for the cases of white, blue, pink, red, and violet noises. As we can observe, for an SNR smaller than −25 dB the decision of segmentation is nearly random using all LLDs. For high SNR values, the segmentation based on E shows better results. The entropy-based LLDs showed a different behavior: when the SNR decreases the result of the segmentation enhances because weak correlations are broken by the addition of random noise of high variance. The maximum points are reached at SNR ≈ 5 ± 5 dB. Among the entropy features, the WPE better captures the amplitude differences in the syllables. In the case of white noise, E outperforms the other descriptors, even in situations with low SNR, except for the pink noise.

Fig. 7 depicts the AUC variation in terms of peak density added to the signals, i.e., the percentage of signal points changed by $\pm\delta$. The performance of all features quickly decreases when the amount of peak noise increases. The exceptions are WPE and H_{FFT}: after the initial point, in approximately 0.05% of peak density, the segmentation improves. The reason is that the addition of the stochastic component helps break the weak correlations. WPE has the worst downward trend due to the increase in the relative frequency of some patterns (π_j) and their high variances.

Table 3
AEER quantifying the error rate of the syllables retrieved. The last line shows the average AEER (or Macro-AEER). The *t*-test with significance level $p \leq 0.05$ was applied to compare E to the rest of LLD. The best values are highlighted in bold. The notation for the different LLD used is the same used in Table 1.

	Features					
Species	E	PE	WPE	PME	ZCR	H_{FFT}
Adenomera h.	0.06	2.08	2.67	2.74	3.32	3.64
Hyla m.	2.18	5.21	4.84	5.15	14.46	1.11
Adenomera a.	0.23	5.44	4.92	5.60	5.13	3.96
Ameerega t.	0.88	0.93	0.97	1.02	2.86	1.02
Osteocephalus o.	2.52	13.66	11.32	14.09	14.24	23.86
Rhinella g.	1.60	0.00	1.00	0.60	28.60	1.40
Scinax r.	1.35	1.78	2.06	2.05	3.46	1.04
Hypsiboas c.	0.12	3.09	3.11	3.44	3.43	1.74
Brachycephalus e.	0.72	0.74	0.97	0.86	1.00	1.00
Aplastodiscus albof.	0.00	6.70	6.26	6.94	6.24	1.03
Aplastodiscus albos.	1.70	18.70	22.29	26.29	20.64	1.82
Aplastodiscus p.	1.55	0.00	0.00	1.03	10.18	1.07
Dendropsophus a.	0.45	6.12	4.47	5.37	5.17	1.10
Dendropsophus e.	0.00	12.06	12.86	18.46	1.40	1.13
Macro-AEER	**0.95**	5.46	5.55	6.69	8.58	**3.21**

6.2. Event-to-event analysis

In Table 1, we have presented the total number of syllables retrieved by each acoustic descriptor being considered. The rows of the table were separated by species to clearly show the difficulty in segmenting the different vocalization patterns. In some cases, such as the Adenomera a. or Osteocephalus o. species, the descriptors often found more syllables than actually exist, caused by micro segmentation (e.g. one syllable is split in two). In the opposite case, e.g. Brachycephalus e., the total number of syllables was lower than the actual values, which indicates that the descriptors were not sensitive enough. Among the columns, E produces is closer to GT.

Analyzing the results by the number of recovered syllables may result misleading. Since each syllable can be considered as an isolated acoustic event, we need to quantify how many events are lost or incorrectly recovered. Given this context, we present the Acoustic Event Error Rate (AEER), a useful metric to quantify the event-to-event errors. This metric is frequently used for audio context-detection (Giannoulis et al., 2013), and it was applied to the bioacoustic segmentation problem by Colonna et al. (2015).

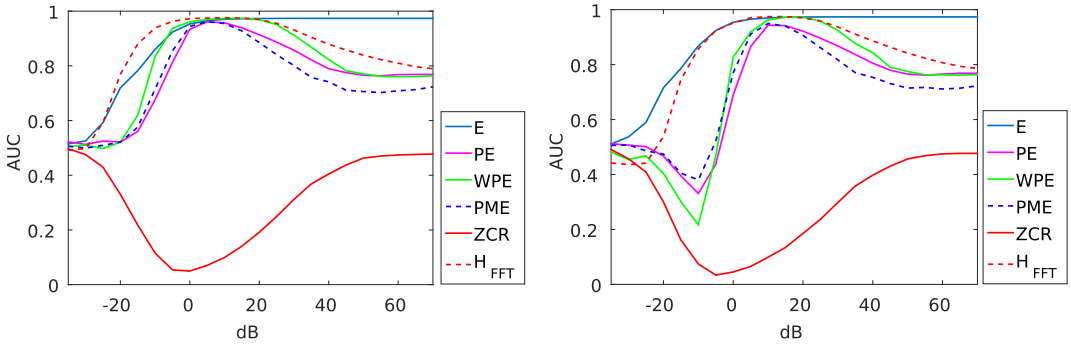
The AEER is defined as:

$$\text{AEER} = \frac{D + I + U}{K}, \tag{18}$$

in which K is the number of syllables in each recorded call, D is the number of missed syllables, I the number of extra syllables, and U the number of replaced syllables computed as $U = \min(D, I)$[3]. We consider that an event is correctly segmented if it starts and ends within ±50 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, AEER = 0.

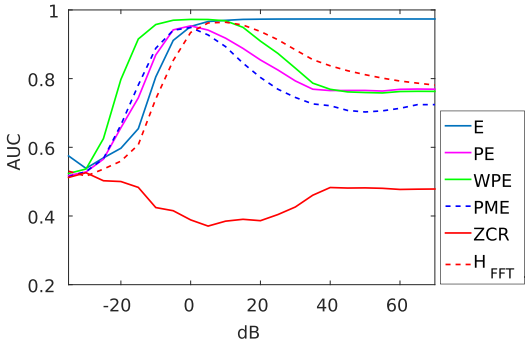
The AEER values presented in Table 3 show that the lowest event-to-event error was achieved by the energy and the spectral entropy. This indicates that fewer events have been lost or mistakenly added. In addition, these two LLDs can retrieve syllables longer than half the original total time. Retrieving syllables without fragmenting them may favor any future classification method.

[3] An implementation of (Giannoulis et al., 2013) can be found at https://goo.gl/6H5Ucm.

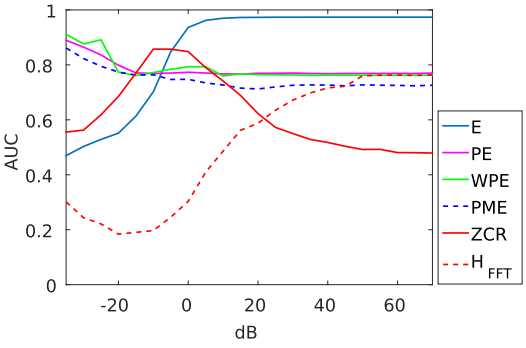


(a) White Noise ($\alpha = 0$)

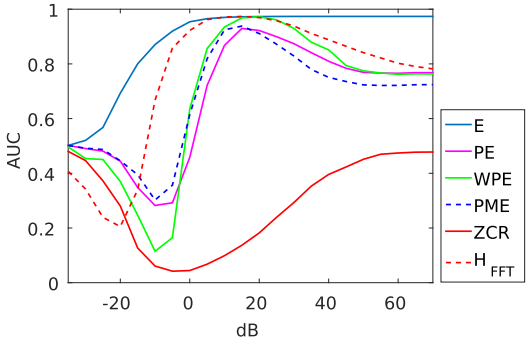
(b) Blue Noise ($\alpha = -1$)



(c) Pink Noise ($\alpha = 1$)



(d) Red Noise ($\alpha = 2$)



(e) Violet Noise ($\alpha = -2$)

Fig. 6. The relationship between the SNR and the AUC variation. These curves show the segmentation performance of each LLD among the extreme levels of SNR. The values at 0 dB are consistent with Table 2. The notation for the different LLD used is the same used in Table 1.

In our formulation, the AEER considers only events omitting the class labels. Therefore, an AEER value can be obtained for each descriptor in each recording. The last line of this table corresponds to the average AEER, also known as Macro-AEER. As a general rule, the Macro-metrics are recommended when the dataset is unbalanced (Sokolova & Lapalme, 2009). Table 3 can be compared with Table 1 to realize the proportion of syllables retrieved and errors incurred.

6.3. Comparing the segmentation using point-to-point metrics

To measure the accuracy of the estimated boundaries, we compared the GT to the automatic segmentation, by counting point-to-point errors. Thus, each point of the segmented signal is compared to each point of the GT segmentation by using metrics based on a decision table (see Section 4.8). 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, the better the automatic segmentation. Table 4 shows Precision, Recall and F-Score of each feature in all records with only the original background noise. We highlighted the results that show statistically significant gains over the best value of each row, considering a 95% confidence level. As we can note, E had the highest precision and F1 values meaning that the binary test correctly identifies the syllable's points. On the other hand, the entropy-based features PE,

Table 4
Precision, Recall and F-Score from point-to-point boundaries evaluation. The bold numbers represent statistically significant ($p \leq 0.05$) differences comparing all LLD to the highest value of each row.

	E	PE	WPE	PME	ZCR	H_{FFT}
Pre	**0.87**	0.39	0.36	0.34	0.12	0.15
Rec	0.53	**0.93**	**0.96**	**0.95**	0.44	0.23
F1	**0.61**	**0.48**	0.46	0.44	0.16	0.13

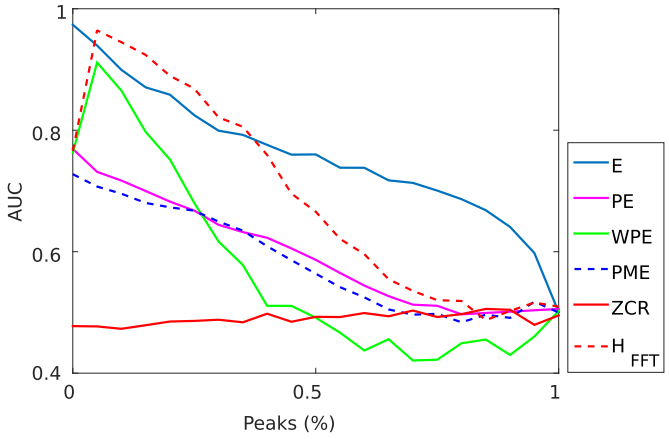


Fig. 7. AUC variation considering the percentage of peak noise. The values at 0% are consistent with the Table 2. The notation for the different LLD used is the same used in Table 1.

WPE and PME had the highest recall meaning that the binary classification test better identifies the negative points. This result is general and is held for the majority of the species recorded.

We used an audio stream of the Aplastodiscus perviridis to depict a visual example of what happens when we add white noise at different SNRs. We choose this example because the background noise of the original signal in this recording is almost uncorrelated. Fig. 8 presents three plots with FNR, FPR, and Accuracy (Acc), considering different SNR (see Eqs. (14)–(16)).

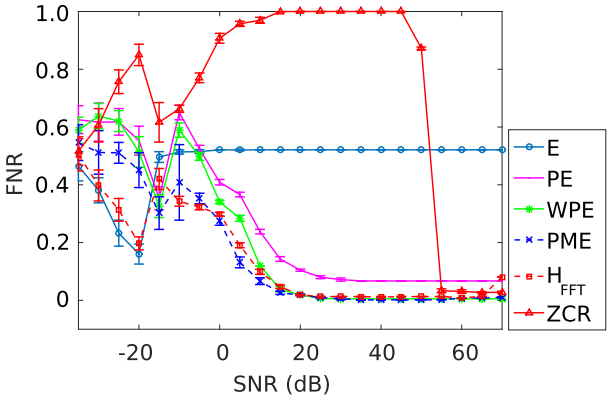
In these plots, we can highlight interesting points. First, the accuracy of entropy features is higher than E. This suggests that the addition of a purely random variable (even with low amplitude) is enough to break the weak correlations of the background noise. Secondly, Fig. 8c shows that E and H_{FFT} are more robust to the increase of σ_ξ, keeping a high accuracy until −15 dB. In addition, the H_{FFT} curve at 45 dB, depicts a change in the behavior, which is caused by the increment in the noise floor, that helps equalize the spectrum and reduce the entropy values.

Similar considerations hold for FPR and FNR, except for E, in Fig. 8a, which has a higher miss rate compared to the entropy LLDs. In this case, the segmentation using E caused the loss of points at the beginning and end of the syllables, indicating that the segmentation criterion is extremely strict.

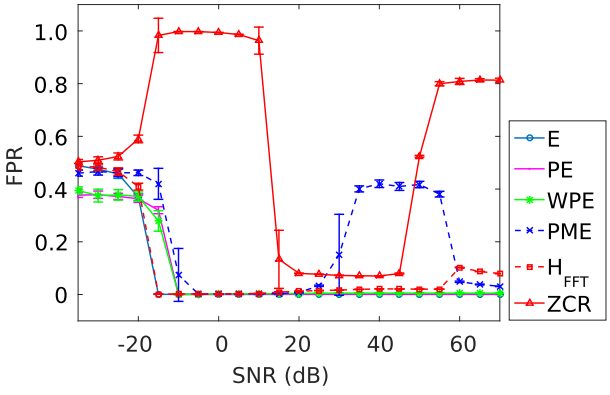
6.4. Ranking and combination of LLDs

Since each acoustic feature is not necessarily related to the others, combining features might improve performance. In order to address this hypothesis, two issues must be considered: (1) how to combine features avoiding the combinatorial problem with exponential complexity, and (2) how to reduce the combination to a 1-dimensional array to apply Algorithm 2. Hence, we decided to rank the LLDs according to the Information Gain criterion and apply Principal Component Analysis (PCA) for dimensional reduction.

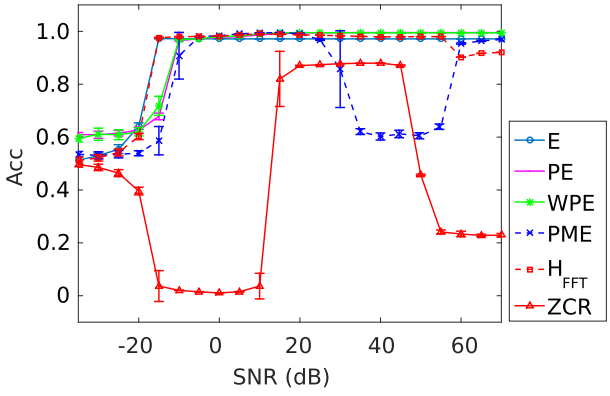
Information Gain (IG) evaluates attributes by measuring the reduction of uncertainty with respect to the class considering the entropy as a measure of "impurity" (Witten & Frank, 2005). In other words, IG measures the impurity reduction caused by each feature in a collection of samples. Thus, features that perfectly partition the set of classes should give maximal information, and unrelated features should give no information. The LLD ranking and their IGs are shown in Table 5 for two different noise conditions: with background noise (**BN**) and white noise at 0 dB. Comparing the IG columns of this table, we realize that an increase in SNR leads to a decrease in IG. Regardless, the AUC values of Table 6 show im-



(a) False Negative Rate or miss rate.



(b) False Positive Rate false alarm.



(c) Accuracy.

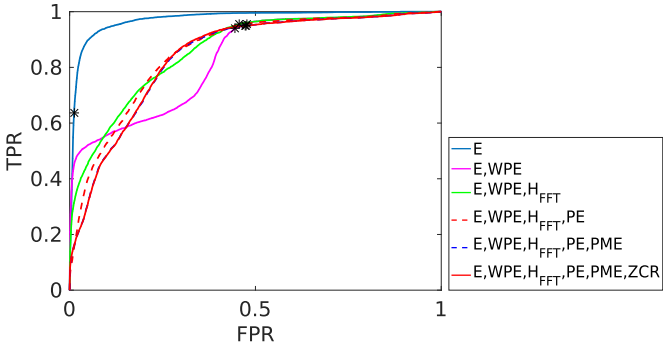
Fig. 8. An example of the behavior of point-to-point metrics FPR and FNR when we add white noise ($\alpha = 0$) and vary the SNR for the species Aplastodiscus Perviridis.

Table 5
Information Gain rankings under normal background noise (left) and white noise at 0 dB (right).

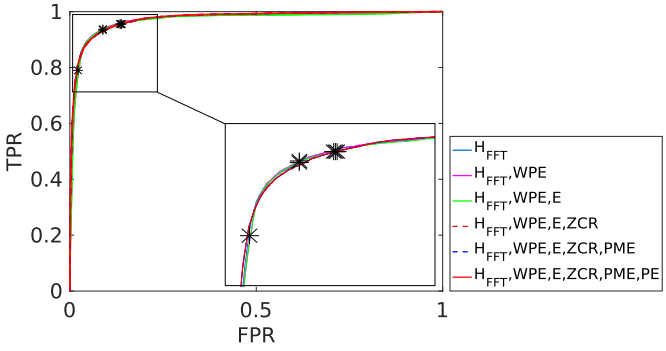
	BN		White noise	
Ranking	LLD	IG	LLD	IG
1st	E	0.4811	H_{FFT}	0.3163
2nd	PE	0.4809	WPE	0.2738
3rd	WPE	0.4807	E	0.2727
4th	H_{FFT}	0.4807	ZCR	0.2643
5th	PME	0.2282	PME	0.2623
6th	ZCR	0.1480	PE	0.1243

Table 6
AUC values of the ROC curves shown in Fig. 9.

BN		White noise	
LLDs	AUC	LLDs	AUC
E	0.973	H_{FFT}	0.972
E,WPE	0.830	H_{FFT}, WPE	0.971
E,WPE,H_{FFT}	0.862	H_{FFT}, WPE,E	0.969
E,WPE,H_{FFT} PE	0.854	H_{FFT}, WPE,E ZCR	0.974
E,WPE,H_{FFT} PE,PME	0.841	H_{FFT}, WPE,E ZCR,PME	0.974
E,WPE,H_{FFT} PE,PME,ZCR	0.842	H_{FFT}, WPE,E ZCR,PME,PE	0.975



(a) **BN**.



(b) White noise (0 dB).

Fig. 9. ROC curves of LLD combination and reduction. The PCA output represent an incremental combination of LLDs. The notation for the different LLD used is the same used in Table 1.

provements for some LLD combinations. This fact confirms once again that the presence of white noise improves the performance of LLDs.

After ranking, the LLDs were combined sequentially and reduced via PCA. Fig. 9a and b show the ROC curves of such combinations applying the same methodology of Section 6.1. The AUC of each curve is shown in Table 6. Among all curves the signal's Energy performs better under **BN** while the H_{FFT} is better in the presence of white noise. The combinations including PE, WPE, and PME achieved similar performance among them. Also, when we add ZCR to the set of combinations under **BN**, the performance reduces even more. We conclude that feature combination in the first scenario (Fig. 9a) does not improve the segmentation performance as we expected. However, in the second scenario (Fig. 9b), with white noise, all combinations obtain a similar performance (see Table 6). Beyond that, and given the variations of AUC, we note that the ranking based on the IG coefficient can be misleading for the LLDs tested here.

Finally, Fig. 10 shows a visual example of segmentation using three features separately and their PCA combination. These features (E, WPE, and H_{FFT}) are essentially unrelated because they were obtained through different procedures. Fig. 10 also shows a mix of syllables belonging to three different species, i.e., we concatenate the three vocalizations of different species in only one stream. The vertical dotted lines mark the start and end point of each vocalization. In addition to the background noise, we added white noise at 20 dB. We can then visually check the boundaries found using each LLD. In this case, E was too strict, because it cuts parts of the syllables or loses the whole syllable (e.g, as the last one). In contrast, the boundaries of WPE and H_{FFT} were less strict incurring in two false positives. The PCA reduction achieved a balanced performance neither strict nor tolerant.

7. Conclusions and final comments

In this work, we presented a comprehensive evaluation of different Low-Level acoustic Descriptors (LLDs or features) used for automatically segmenting anuran calls. As an additional contribution, we showed that, depending on the noise pattern, the Permutation Entropy (PE) quantifier and its variants can improve signal segmentation. The idea is to combine the simplicity of the energy models with the robustness of the probabilistic models in an unsupervised manner. Hence, we computed the entropy value of the PDF derived from the signal and used it as an LLD. Considering the frame size n, both features the Energy and ZCR have linear computational complexity $\Theta(n)$; the H_{FFT}, which depends on the FFT transform, has a complexity $\Omega(n\log_2(n))$ in the best case and $O(n^2)$ in the worst case. The LLDs derived from the PE methodology have $\Omega(n)$ and $O(mn)$ complexities, for the best and the worst cases, respectively, in which m is the embedding dimension. In addition, we presented an algorithm to find the optimal segmentation threshold (Algorithm 2).

We showed that for signals with abrupt amplitude changes, uncorrelated background noise, and low SNR, the entropy based on the WPE methodology is the best option, unless we have more than 0.05% peaks of impulsive noise. This LLD has a linear computational complexity, weighted by the constant m (embedding dimension), being higher than Energy and lower than H_{FFT}. For the cases in which the noise is completely white, and the spectrum of the noise floor is approximately uniform, the H_{FFT} is the best choice. However, the computational complexity of the FFT transform should be considered if the final application runs in a resource-constrained sensor.

Permutation Entropy and its variants perform well with pink and red noise (low frequencies), but they experiment lower accuracy with blue and violet noise (high frequencies). Moreover, the addition of white noise improves the segmentation for almost all entropy-based LLDs, because it breaks weak correlations. This effect was also reported by Bandt and Pompe (2002) and Veisi, Pariz, and Karimpour (2007), being a consequence of the invariant transformation property of PE. In the case of spectral entropy, the addition of white noise improves the segmentation, because this noise spreads energy uniformly in all frequencies, masking the small peaks of the spectrogram. Most likely, the spectrum is more concentrated around the fundamental frequencies, hence, decreasing the entropy.

By considering every syllable as an acoustic event and measuring the Acoustic Event Error Rate (AEER), we observed that most of the LLDs produce several false positive samples of events (micro-segments) that might affect the species (syllable) recognition step. One reason is the length of the intervals between each syllable, which is closely related to the frame size chosen. The frame size and the overlap factor were not varied in our experiments and can be done in a future evaluation. Also, a postprocessing method, such

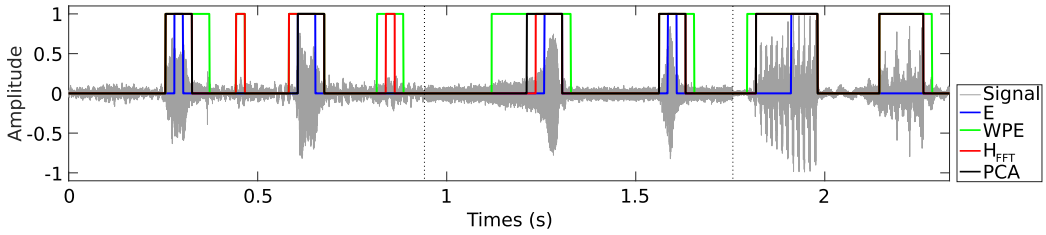


Fig. 10. Example of segmentation boundaries in a stream with three different species: Adenomera h., Hyla m., and Scinax r., contaminated with white noise at 20 dB. The dotted vertical lines indicate the beginning and end of each species.

as smoothing filter, should be evaluated in order to eliminate a greater number of false positives. However, we keep the method as simple as possible, avoiding any extra processing block, unless it is extremely necessary.

The overall accuracy of our method depends on the threshold selected during the segmentation. We presented evidences, such as the position of the black points in Fig. 5 or the quality of the boundaries in Fig. 10, suggesting that Algorithm 2 has optimal performance for the given task. We would like to emphasize that ROC curves represent the output of all possible thresholds regardless the technique adopted to find them. Besides that, a continuous audio stream coming from a real scenario can present a non-stationary behavior of background noise, in which case, the threshold should be updated regularly. The threshold updates can be performed onboard the sensor node, without any human intervention or data transmission/reception, just by buffering the feature values corresponding to the most recent frames and recalculating the optimal threshold. Such update would be infeasible with a supervised technique.

We would like to emphasize that for a practical deployment in a sensor node, a initial setup stage must be included, in which the first signal frames should be used to obtain the optimal threshold. Otherwise, if the goal is to segment a call previously stored in a database, we can use all the signal frames to find the optimal segmentation. As a grouping method, there is no need to label data, we should only ensure that sample signal and background noise are available when the threshold is calculated. Nevertheless, our results provide a lower bound for any future segmentation technique that uses information theory.

In a future work, Algorithm 2 can be transformed to an adaptive method by replacing lines 4 and 5 for the incremental mean calculation (Finch, 2009). With the incremental mean calculation the threshold value T (line 7 of same algorithm) can be updated using each incoming frame of the input signal.

As a general conclusion, we recommend that before choosing any LLD one should test the type of noise and the level of the noise floor. Furthermore, it is important to avoid the zero crossing rate (ZCR), because it is highly sensitive to the noise present in rain forests. Moreover, the type of the rain forest noise is a matter rarely studied in the related works. A more accurate noise model could help do a more realistic simulation in future work. Another interesting future work is the implementation and evaluation of a signal filter before the segmentation step.

Acknowledgments

Juan Colonna acknowledges to National Council of Technological and Scientific Development (CNPq, Brazil), FAPEAM (PROTI) and CAPES for the PhD scholarship. Eduardo Nakamura acknowledges to FAPEAM for the support granted through the Anura Project (FAPEAM/CNPq PRONEX 023/2009) and CNPq (process 309471/2015-0). Osvaldo A. Rosso acknowledges the financial support from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

References

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. *PeerJ, 1*(e103), 1–19.

Alonso, J. B., Cabrera, J., Shyamnani, R., Travieso, C. M., Bolaños, F., García, A., et al. (2017). Automatic anuran identification using noise removal and audio activity detection. *Expert Systems with Applications, 72*(Supplement C), 83–92. doi:10.1016/j.eswa.2016.12.019.

Ballón, M., Bertrand, A., Lebourges-Dhaussy, A., Gutiérrez, M., Ayón, P., Grados, D., et al. (2011). Is there enough zooplankton to feed forage fish populations off peru? an acoustic (positive) answer. *Progress in Oceanography, 91*(4), 360–381.

Bandt, C., & Pompe, B. (2002). Permutation entropy: A natural complexity measure for time series. *Physical Review Letters, 88*(17), 1–5.

Bardeli, R. (2009). Similarity search in animal sound databases. *IEEE Transactions on Multimedia, 11*(1), 68–76.

Brandes, T. S. (2008). Feature vector selection and use with hidden Markov models to identify frequency-modulated bioacoustic signals amidst noise. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(6), 1173–1180.

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*(4), 903–913.

Cettolo, M., Vescovi, M., & Rizzi, R. (2005). Evaluation of bic-based algorithms for audio segmentation. *Computer Speech & Language, 19*(2), 147–170.

Cheng, S.-S., & Wang, H.-M. (2003). A sequential metric-based audio segmentation method via the Bayesian information criterion. In *Proceedings of the European conference on speech communication and technology (INTERSPEECH)* (pp. 945–948).

Chu, W., & Blumstein, D. T. (2011). Noise robust bird song detection using syllable pattern-based hidden Markov models. In *Proceedings of the international conference on acoustics, speech and signal processing (ICASSP)* (pp. 345–348).

Cole, E. M., Bustamante, M. R., Reinoso, D. A., & Funk, W. C. (2014). Spatial and temporal variation in population dynamics of andean frogs: Effects of forest disturbance and evidence for declines. *Global Ecology and Conservation, 1*(0), 60–70.

Colonna, J. G., Cristo, M. A. P., & Nakamura, E. F. (2014). A distribute approach for classifying anuran species based on their calls. In *Proceedings of the 22nd international conference on pattern recognition (ICPR)* (pp. 1242–1247).

Colonna, J. G., Cristo, M. A. P., Salvatierra, M., & Nakamura, E. F. (2015). An incremental technique for real-time bioacoustic signal segmentation. *Expert Systems with Applications, 42*(21), 7367–7374.

Colonna, J. G., Peet, T., Ferreira, C. A., Jorge, A. M., Gomes, E. F., & Gama, J. (2016). Automatic classification of anuran sounds using convolutional neural networks. In *Proceedings of the ninth international c* conference on computer science & software engineering* (pp. 73–78).

Colonna, J. G., Ribas, A. D., Santos, E. M. d., & Nakamura, E. F. (2012). Feature subset selection for automatically classifying anuran calls using sensor networks. In *Proceedings of the international joint conference on neural networks (JCNN)* (pp. 1–8).

Depraetere, M., Pavoine, S., Jiguet, F., Gasc, A., Duvail, S., & Sueur, J. (2012). Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland. *Ecological Indicators, 13*(1), 46–54.

Evangelista, T. L. F., Priolli, T. M., Silla, C. N., Angelico, B. A., & Kaestner, C. A. A. (2014). Automatic segmentation of audio signals for bird species identification. In *Proceedings of the international symposium on multimedia (ISM)* (pp. 223–228).

Fadlallah, B., Chen, B., Keil, A., & Príncipe, J. (2013). Weighted-permutation entropy: A complexity measure for time series incorporating amplitude information. *Physical Review E, 87*(2), 1–7.

Fagerlund, S., & Laine, U. K. (2014). Classification of audio events using permutation transformation. *Applied Acoustics, 83*(1), 57–63.

Fawcett, T. (2006). An introduction to Roc analysis. *Pattern Recognition Letters, 27*(8), 861–874.

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.

In *Proceedings of the international conference on multimedia and expo (ICME)* (pp. 452–455).
Giannakopoulos, T., Pikrakis, A., & Theodoridis, S. (2008). A novel efficient approach for audio segmentation. In *Proceedings of the nineteenth 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 *Proceedings of the European signal processing conference (EUSIPCO)* (pp. 1–5).
Gibbs, J. P., Whiteleather, K. K., & Schueler, F. W. (2005). Changes in frog and toad populations over 30 years in new york state. *Ecological Applications, 15*(4), 1148–1157.
Heinicke, S., Kalan, A. K., Wagner, O. J. J., Mundry, R., Lukashevich, H., & Kuhl, H. S. (2015). Assessing the performance of a semi-automated acoustic monitoring system for primates. *Methods in Ecology and Evolution, 6*(7), 753–763.
Huang, C. J., Yang, Y. J., Yang, D. X., & Chen, Y. J. (2009). Frog classification using machine learning techniques. *Expert Systems with Applications, 36*(2), 3737–3743.
Jaafar, H., & Ramli, D. A. (2013). Automatic syllables segmentation for frog identification system. In *Proceedings of the ninth international colloquium on signal processing and its applications (CSPA)* (pp. 224–228).
Jaafar, H., Ramli, D. A., & Shahrudin, S. (2013). MFCC based frog identification system in noisy environment. In *Proceedings of the international conference on signal and image processing applications (ICSIPA)* (pp. 123–127).
Kamper, H., Livescu, K., & Goldwater, S. (2017). An embedded segmental k-means model for unsupervised segmentation and clustering of speech. Computing Research Repository, (arXiv) abs/1703.08135.
Kasdin, N. J. (1995). Discrete simulation of colored noise and stochastic processes and $\frac{1}{|f|^{\alpha}}$ power law noise generation. *Proceedings of the IEEE, 83*(5), 802–827.
Labate, D., Foresta, F. L., Morabito, G., Palamara, I., & Morabito, F. C. (2013). Entropic measures of eeg complexity in Alzheimer's disease through a multivariate multiscale approach. *IEEE Sensors Journal, 13*(9), 3284–3292.
Lopes, M. T., Koerich, A. L., Silla, C. N., & Kaestner, C. A. A. (2011). Feature set comparison for automatic bird species identification. In *Proceedings of the international conference on systems, man, and cybernetics (SMC)* (pp. 965–970).
Lowen, S. B., & Teich, M. C. (1990). Power-law shot noise. *IEEE Transactions on Information Theory, 36*(6), 1302–1318.
Luque, A., Romero-Lemos, J., Carrasco, A., & Barbancho, J. (2017). Non-sequential automatic classification of anuran sounds for the estimation of climate-change indicators. *Expert Systems with Applications, 95*. doi:10.1016/j.eswa.2017.11.016.
MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology, 84*(8), 2200–2207.
Mcllraith, A. L., & Card, H. C. (1997). Birdsong recognition using backpropagation and multivariate statistics. *IEEE Transactions on Signal Processing, 45*(11), 2740–2748.
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*(1), 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*(3), 1–55.
Neal, L., Briggs, F., Raich, R., & Fern, X. Z. (2011). Time-frequency segmentation of bird song in noisy acoustic environments. In *Proceedings of the international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2012–2015).
Noda, J. J., Travieso, C. M., & Sánchez-Rodríguez, D. (2016). Methodology for automatic bioacoustic classification of anurans based on feature fusion. *Expert Systems with Applications, 50*(Supplement C), 100–106. doi:10.1016/j.eswa.2015.12.020.
Oliveira, A. G., Ventura, T. M., Ganchev, T. D., Figueiredo, J. M., Jahn, O., Marques, M. I., et al. (2015). Bird acoustic activity detection based on morphological filtering of the spectrogram. *Applied Acoustics, 98*(1), 34–42.
Plaszczynski, S. (2007). Generating long streams of $\frac{1}{f^{\alpha}}$ noise. *Fluctuation and Noise Letters, 07*(01), 1–13.
Potamitis, I. (2014). Automatic classification of a taxon-rich community recorded in the wild. *PLOS ONE, 9*(5), 1–11.
Rahman, M., & Bhuiyan, A. (2012). Continuous bangla speech segmentation using shortterm speech features extraction approaches. *International Journal of Advanced Computer Sciences and Applications, 3*(1), 11.
Ren, Y., Johnson, M. T., Clemins, P. J., Darre, M., Glaeser, S. S., Osiejuk, T. S., et al. (2009). A framework for bioacoustic vocalization analysis using hidden Markov models. *Algorithms, 2*(4), 1410–1428.
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 *Proceedings of the international joint conference on neural networks (JCNN)* (pp. 1–7).
Rosso, O. A., Larrondo, H. A., Martin, M. T., Plastino, A., & Fuentes, M. A. (2007). Distinguishing noise from chaos. *Physical Review Letters, 99*(15), 1–4.
Rudnick, D. L., & Davis, R. E. (2003). Red noise and regime shifts. *Deep Sea Research Part I: Oceanographic Research Papers, 50*(6), 691–699.
Sarkar, A., & Sreenivas, T. V. (2005). Automatic speech segmentation using average level crossing rate information. In *Proceedings of the international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 397–400).
Sezgin, M., & Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging, 13*(1), 146–168.
Shen, J., Hung, J., & Lee, L. (1998). Robust entropy-based endpoint detection for speech recognition in noisy environments. In *Proceedings of the fifth international conference on spoken language processing (ICSLP)* (pp. 232–235).
Sinn, M., Keller, K., & Chen, B. (2013). Segmentation and classification of time series using ordinal pattern distributions. *The European Physical Journal Special Topics, 222*(2), 587–598.
Slaby, A. (2007). ROC analysis with MATLAB. In *Proceedings of the twenty ninth international conference on information technology interfaces (ITI)* (pp. 191–196).
Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437.
Somervuo, P., Harma, A., & Fagerlund, S. (2006). Parametric representations of bird sounds for automatic species recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(6), 2252–2263.
Soriano, M. C., Zunino, L., Rosso, O. A., Fischer, I., & Mirasso, C. R. (2011). Time scales of a chaotic semiconductor laser with optical feedback under the lens of a permutation information analysis. *IEEE Journal of Quantum Electronics, 47*(2), 252–261.
Strout, J., Rogan, B., Seyednezhad, S. M. M., Smart, K., Bush, M., & Ribeiro, E. (2017). Anuran call classification with deep learning. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2662–2665). doi:10.1109/ICASSP.2017.7952639.
Sueur, J., Pavoine, S., Hamerlynck, O., & Duvail, S. (2008). Rapid acoustic survey for biodiversity appraisal. *PLOS ONE, 3*(12), 1–9.
Theodorou, T., Mporas, I., & Fakotakis, N. (2014). An overview of automatic audio segmentation. *International Journal of Information Technology and Computer Science (IJITCS), 6*(11), 1.
Tomasini, M., Smart, K., Menezes, R., Bush, M., & Ribeiro, E. (2017). Automated robust anuran classification by extracting elliptical feature pairs from audio spectrograms. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2517–2521). doi:10.1109/ICASSP.2017.7952610.
Vasseur, D. A., & Yodzis, P. (2004). The color of environmental noise. *Ecology, 85*(4), 1146–1152.
Veisi, I., Pariz, N., & Karimpour, A. (2007). Fast and robust detection of epilepsy in noisy eeg signals using permutation entropy. In *Proceedings of the seventh international symposium on bioinformatics and bioengineering* (pp. 200–203).
Voss, R. F., & Clarke, J. (1978). "1/F noise" in music: Music from 1/f noise. *The Journal of the Acoustical Society of America, 63*(1), 258–263.
Wichern, G., Xue, J., Thornburg, H., Mechtley, B., & Spanias, A. (2010). Segmentation, indexing, and retrieval for environmental and natural sounds. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(3), 688–707.
Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd). Morgan Kaufmann.
Wu, B.-F., & Wang, K.-C. (2005). Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. *IEEE Transactions on Speech and Audio Processing, 13*(5), 762–775.
Xie, J., Towsey, M., Truskinger, A., Eichinski, P., Zhang, J., & Roe, P. (2015). Acoustic classification of australian anurans using syllable features. In *Proceedings of the tenth international conference on intelligent sensors, sensor networks and information processing (ISSNIP)* (pp. 2–7).
Yuan, X., Martínez, J.-F., Eckert, M., & López-Santidrián, L. (2016). An improved Otsu threshold segmentation method for underwater simultaneous localization and mapping-based navigation. *Sensors, 16*(7), 1148.
Zunino, L., Olivares, F., & Rosso, O. A. (2015). Permutation min-entropy: An improved quantifier for unveiling subtle temporal correlations. *Europhysics Letters (EPL), 109*(1), 1–6.
Zunino, L., Soriano, M. C., & Rosso, O. A. (2012). Distinguishing chaotic and stochastic dynamics from time series by using a multiscale symbolic approach. *Physical Review E, 86*(4), 1–10.