Threshold Estimation in Energy-Based Methods for Segmenting Birdsong Recordings

José Francisco Ruiz-Muñoz[1], Mauricio Orozco-Alzate[1,2], and César Germán Castellanos-Domínguez[1]

[1] Grupo de Procesamiento y Reconocimiento de Señales, Universidad Nacional de Colombia - Sede Manizales, km 7 vía al aeropuerto, Manizales (Caldas), Colombia
[2] Departamento de Informática y Computación, Universidad Nacional de Colombia - Sede Manizales, km 7 vía al aeropuerto, Manizales (Caldas), Colombia
{jfruizmu,morozcoa,cgcastellanosd}@unal.edu.co

Abstract. Monitoring wildlife populations is important to assess ecosystem health, attend environmental protection activities and undertake research studies about ecology. However, the traditional techniques are temporally and spatially limited; in order to extract information quickly and accurately about the current state of the environment, processing and recognition of acoustic signals are used. In the literature, several research studies about automatic classification of species through their vocalizations are found; however, in many of them the segmentation carried out in the preprocessing stage is briefly mentioned and, therefore, it is difficult to be reproduced by other researchers. This paper is specifically focused on detection of regions of interest in the audio recordings. A methodology for threshold estimation in segmentation techniques based on energy of a frequency band of a birdsong recording is described. Experiments were carried out using chunks taken from the RMBL-Robin database; results showed that a good performance of segmentation can be obtained by computing a threshold as a linear function where the independent variable is the estimated noise.

Keywords: Audio signal processing and recognition, segmentation, bioacoustics.

1 Introduction

Technology for automatic classification of animal vocalizations is a useful tool in research studies on taxonomy, ecology and conservation as well as for attending activities of environmental monitoring. Traditional technologies for assessing ecosystem health, such as line transects or fixed-radius point counts, are spatial and time consuming, often imply expensive and exhausting journeys, and could be disruptive to the habitat under observation; thereby, in order to avoid those inconveniences, an automated system would be desired [1]. Particularly, researchers are interested in analyzing birdsongs because birds are widely distributed in nature, relatively easy to detect by their vocalizations and they have

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 480–487, 2013.

great knowledge of the biology of most species. Furthermore, there is a commercial interest in developing this type of systems due to the large and increasing number of birdwatchers worldwide [2]; additionally, this technology is ideal for impact studies and environmental management plans that are frequently required by authorities in many countries.

The problem of classifying bird species from an audio recording is a typical signal recognition problem [3], therefore in many studies, stages of signal preprocessing, feature extraction and classification are included. The first one includes the segmentation of vocalizations into smaller recognition units [4]. Sometimes it is done manually, and sometimes automatically; nevertheless, automatic recognition should not require manual segmentation. Even if classification is the aim in some papers, it is mentioned that a complete automatic recognition system should include the automatic detection of intervals of interest [5] because accurate segmentation is fundamental for successful classification systems [3].

Segmentation algorithms have been developed using energy and entropy as criteria to identify the onset and offset times of the regions of interest [5–7]. This process is simple in ideal conditions [3]; if vocalization call is the only sound in the recording, an increase in energy reveals a region of interest. In real conditions, the signal is degraded due to the many sources of sound in a recording, e.g., wind streams, background noise from other animals and surrounding events.

In this paper, it is proposed a non-supervised segmentation method of birdsong recordings. The signal energy is calculated from a frequency band extracted from the Short-time Fourier transform (STFT) —it can be considered as an image known as spectrogram: a representation of the intensity of a sound as a function of time and frequency [2,6]. The output of the method is a binary signal in function of time, where time instants of interest are marked with "one" and non-interest time instants with "zero". Despite segmentation methods are essential in many studies about this topic, it is not clear how similar their outputs are to manual segmentations. Such a judgement must be based on objective performance estimation measurements; a number of them are presented in Section 2.4. Three ways of threshold estimation are compared: optimal in each case, linear function of estimated noise (parameter computed by least squares) and the segmentation technique used in [4].

2 Material and Methods

The proposed segmentation method is shown in Fig. 1. Basically, it consists in detecting regions with the highest energies in frequencies where the sound of interest typically exhibits its components.

2.1 Time-Frequency Analysis

Time-frequency analysis of a signal can be carried out through the STFT, namely, the Fourier transform per frame of a signal. STFT is a representation of the distribution of acoustic energy across frequencies and over time. Often,

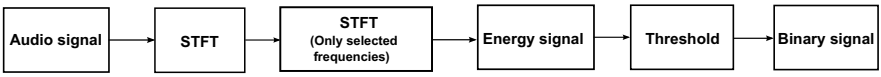

Fig. 1. Flow diagram of the proposed segmentation method

the time is graphically represented in the horizontal axis and frequency in the vertical one, and the amount of power detected is represented as the intensity at each time-frequency point, as follows:

$$\mathbf{S} = [s_{ij}]_{N\times M} = \begin{bmatrix} P_0(f_0) & P_1(f_0) & \cdots & P_{M-1}(f_0) \\ P_0(f_1) & P_1(f_1) & \cdots & P_{M-1}(f_1) \\ P_0(f_2) & P_1(f_2) & \cdots & P_{M-1}(f_2) \\ \vdots & \vdots & \ddots & \vdots \\ P_0(f_{N-1}) & P_1(f_{N-1}) & \cdots & P_{M-1}(f_{N-1}) \end{bmatrix} \tag{1}$$

where $i = 0, 1, ..., N-1$ corresponds to the frequency indexes and $j = 0, 1, ..., M-1$ to the time.

2.2 Energy Signal

A smoothed energy signal is computed taking only frequencies selected in the spectrogram (it requires *a priori* knowledge). When the segmentation is done in a limited frequency band, with minimum frequency f_l and maximum frequency f_h, a sub-matrix $\hat{\mathbf{S}} = [\hat{s}_{ij}]_{\hat{N}\times M}$ ($\hat{N} \leq N$) is taken from the STFT representation, such that elements in $\mathbf{S}$ corresponding to $f < f_l$ and $f > f_h$ are discarded. Energy signal is computed as:

$$E_j = \frac{1}{\hat{N}} \sum_{i=0}^{i=\hat{N}-1} \hat{s}_{ij}^2, \tag{2}$$

where the energy vector from (3) is obtained.

$$\mathbf{E} = [E_0,\ E_1,\ ...,\ E_{M-1}] \tag{3}$$

A smoothed energy signal $\hat{\mathbf{E}} = [\hat{E}_0\ \hat{E}_1\ ...\ \hat{E}_{M-1}]$ is obtained using the convolution operator and a Hann window $\mathbf{w} = \frac{1}{\sum_{i=0}^{l-1} w_i}[w_0\ w_1\ ...\ w_{l-1}]$ of size l, where $w_i = 0.54 - 0.46\cos(2\pi i/(l-1))$ to $i = 0, ..., l-1$:

$$\hat{E}_j = (\mathbf{E} * \mathbf{w})[j] = \sum_{i=0}^{i=l-1} E_n \hat{w}_i, \tag{4}$$

with $j = 0, 1, ..., M-1$.

The normalized energy signal is:

$$\hat{\mathbf{E}}_{norm} = 10 \log \left(\frac{\hat{\mathbf{E}}}{\text{argmax}\{\hat{\mathbf{E}}\}} \right), \tag{5}$$

so that the maximum element of $\hat{\mathbf{E}}_{norm}$ is equal to zero (0).

2.3 Binary Signal

Assuming that vocalizations are present in regions where the energy signal $\hat{\mathbf{E}}_{norm}$ has the highest values, a threshold T_{dB} is used to build a binary function $\mathbf{B} = [B_0,\ B_1,\ ...,\ B_{M-1}]$ so:

$$B_i = \begin{cases} 1 & \text{if } \hat{E}_{norm\ i} > T_{dB} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

for $i = 0, ..., M-1$.

2.4 Performance

Let $T = [t_0,\ t_1,\ ...,\ t_{M-1}]$ be a vector of time instants and $\mathbf{A} = [A_0,\ A_1,\ ...,\ A_{M-1}]$ a reference binary signal manually labeled, where $A_i = 1$ if the time instant t_i in the recording is considered as a point of interest and $A_i = 0$ otherwise; and $\mathbf{B} = [B_0,\ B_1,\ ...,\ B_{M-1}]$ is a binary signal resulting from a segmentation process as is indicated in Section 2.3. Recall rate (R) and precision rate (P) were chosen to measure the performance, following the same evaluation protocol used in [8]; the first one relates the number of points of interest correctly detected or hits (N_h, it is the total of points where $A_i = B_i = 1$) with the number of real points of interest (N_r, it is the total of points where $A_i = 1$), according with a manual segmentation; and the second one relates N_h with the number of detected points (N_d, it is the total of points where $B_i = 1$), so:

$$\text{R} = \frac{N_h}{N_r} \times 100\%, \quad \text{P} = \frac{N_h}{N_d} \times 100\%. \tag{7}$$

A measure that combines the two previous ones is the Euclidean distance between the point (R,P) in Cartesian coordinates and the point (100,100):

$$d_T(R, P) = \sqrt{(100 - R)^2 + (100 - P)^2}, \tag{8}$$

where the best segmentation corresponds to 0 and the worst one to 100.

Other measure that combines R and P is the F1 Score, it is defined as:

$$\text{F1 Score} = \frac{2\text{PR}}{\text{P+R}} \tag{9}$$

2.5 Estimating the Optimum Threshold

Threshold level is important because it is used to mark the boundaries between chosen and discarded segments. In [4] it is suggested to choose as threshold half of the noise level N_{dB} (computed by an iterative method, see Section 3) of the energy signal normalized with the maximum value 0 dB.

Intuitively, it is expected that when the T_{dB} value is changed in (6) from N_{dB} until 0 dB, R starts in 100 and tends to 0 and P starts in 0 and tends to 100 (assuming that at least the maximum energy point is considered of interest). Therefore, the best T_{dB} is considered the one that minimizes (8), as follows: if $\mathbf{T}_{dB} = [T_{dB\ 0},\ T_{dB\ 1},\ ...,\ T_{dB\ k}]$, where $T_{dB\ 0} = 0$ and $T_{dB\ k} = N_{dB}$, and the corresponding $\mathbf{R} = [R_0,\ R_1,\ ...,\ R_k]$ and $\mathbf{P} = [P_0,\ P_1,\ ...,\ P_k]$, the optimal threshold (T_{opt}) is $T_{dB\ i}$ for which $d_T(R_k, P_k)$ is minimum.

3 Experimental Setup

The objective in this section is to find a rule to estimate a threshold in order to obtain good segmentation. Experiments were done by following the steps listed below:

- Signal energy was estimated from audio recording chunks with the following features: STFTs representations computed using 512 points in each block of time with an overlap of 256 points; a sub-matrix was estimated, as described in Section 2.2, with $f_l = 1000$ Hz and $f_h = 5000$ Hz because the pitch information of the Robin ranges from 1500 to 4500 Hz [8]. The normalized and smoothed energy signal (see (5)) was estimated using a Hann window of size 20.
- N_{dB} is computed: the initial N_{dB} is set to the lowest $\hat{\mathbf{E}}_{norm}$ level and updated as the mean from gaps between regions where $\hat{\mathbf{E}}_{norm} < N_{dB}/2$ until the previous and current values of N_{dB} not vary more than 1 dB.
- T_{opt} is estimated as it is explained in Section 2.5: segmentation was carried out to several levels of threshold $\mathbf{T}_{dB}$, 20 steps from 0 to the minimum of $\hat{\mathbf{E}}_{norm}$. The element of $\mathbf{T}_{dB}$ with the best performance is chosen as T_{opt}.
- Threshold in function of N_{dB} is computed using least squares: this method consists in minimizing the expression $\left\| \hat{y} - [\hat{x}\ \ 1]\,[m\ \ c]^T \right\|^2$, where $\hat{x}$ and $\hat{y}$ are vectors that represents points in a Cartesian coordinate system, and m and c are the parameters of the linear equation $f(x) = mx + c$ computed in the regression; N_{dB} and T_{opt} for each chunk were taken as $\hat{x}$ and $\hat{y}$ respectively. Implementation was done using the numerical package of Python "Numpy"; the command `numpy.linalg.lstsq` was used.

3.1 Dataset

Experiments were carried out using a set of chunks of the RMBL-Robin database, which can be downloaded from `http://www.ee.ucla.edu/~weichu/bird/`. It is

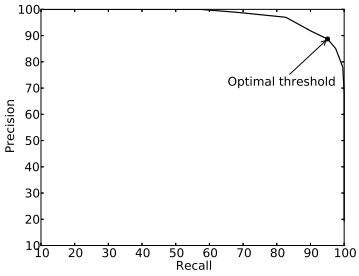

Fig. 2. P-R curve from a segment of recording

a set of recordings of Robin bird songs collected at the Rocky Mountain Biological Laboratory near Crested Butte, Colorado (USA) in the summer of 2009 [8]. Recordings are corrupted by environmental background noises and human voices. This dataset was selected in this research because timing boundaries of syllables were annotated by experts. Although noise and intensity levels vary considerably between recordings, it was assumed that they are relatively constant over chunks of approximately 10 seconds of duration as considered in [6]. Therefore, experiments were carried out using 20 segments of 10 s approximately, extracted from 10 recordings. Names of the selected recordings and time intervals are specified in Table 1.

Table 1. Chunks selected from the RMBL-Robin database

Name of file .wav	Start time	End time
A-01june09-0702-robin.wav	0	10
A-30may09-0729-robin.wav	24	36
C-30may09-0826-robin.wav	0	12
C-31may09-0608-robin.wav	9	19
E-01june09-0537-robin.wav	0	12
E-01june09-0543-robin.wav	0	9
G-08june09-0517-robin.wav	0	12
H-08june09-0507-robin.wav	12	23
H-08june09-0512-robin.wav	0	10
H-09june09-0518-robin.wav	14	26
L-03june09-1905-robin.wav	9	21
L-12june09-0728-robin.wav	0	11
S-09june09-1953-robin.wav	2	15
U-03june09-1813-robin.wav	4	16
U-03june09-1815-robin.wav	10	24
W-03june09-1905-robin.wav	0	9
W-08june09-0733-robin.wav	49	63
W-09june09-1902-robin.wav	4	15
X-04june09-0615-robin.wav	5	15
X-14june09-0518-robin.wav	6	24

3.2 Results

Table 2 shows the performance obtained in segmentation experiments. Two methods of threshold estimation were compared with the performance obtained

for segmentation with T_{opt} in each chunk: **1) fitted line:** threshold in function of N_{dB} ($T_{dB} = m\ N_{dB} + c$), where m and c are obtained with the regression as explained in Section 3; **2) half of noise:** the same as the previous case but with $m = 0.5$ and $c = 0$, as proposed in [4] (see Fig. 3).

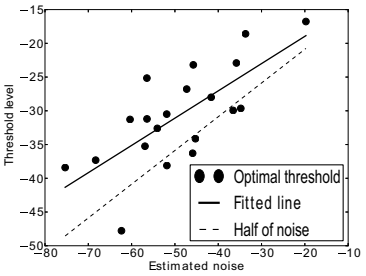

Fig. 3. Graphical representation of optimal threshold estimated in each chunk and the linear functions obtained by the linear regression (fitted line) and taken the threshold equals to $N_{dB}/2$ (half of noise)

Notice in Table 2 that, according to d_T and F1 Score measures, with threshold regression it is obtained a better performance than the one using the method of [4] (half of N_{dB}). With threshold estimation as half of N_{dB} a good P was obtained, even better than with optimum threshold by chunk, nevertheless P decays considerably.

Table 2. Comparison between performance obtained by computing the threshold as a linear function of N_{dB}, and the optimal threshold estimated by each chunk. If the performance measure is next to ↑: the better the bigger; analogously, if it is next to ↓: the better the lower.

	R ↑	P ↑	d_T ↓	F1 Score ↑	Parameters estimated
Fitted line	**90.3**	80.7	**21.3**	**85.2**	m=0.4; c=-10.9
Half of N_{dB}	76.0	**89.8**	26.1	82.3	m=0.5; c=0
Optimum threshold (Reference)	91.1	85.4	17.1	88.2	

4 Conclusions

A detailed segmentation methodology for birdsong recordings was presented. It is based on energy of a frequency band, is straightforward and provides a good performance. The proposed threshold estimation technique consists in computing it in function of background noise from a linear regression. Performance methods are described and results are compared with the optimum threshold heuristically computed for each chunk and with the threshold estimation method from [4].

Several research studies about automated species recognition are not rigorous in the description of the segmentation stage, even when they clarify that their

methods work well when the classification objects are correctly detected [5,9,10]. However, the system confidence depends on both event detection and classification algorithms; therefore, the two stages should be explained in detail, including appropriate evidence of performance.

Birdsongs often have a gramatical structure, where the basic building blocks are called syllables [2] and which have been used as recognition objects in many studies; as future work it is proposed to use a merge and delete criterion particularly to detect these units. Furthermore, other representations different to the energy might be explored, e.g., the entropy and other criteria to detect regions of interest; tuning of the new parameters might be carried out based on the performance measures used in this research.

Acknowledgments. This research was supported by "Programa Nacional de Formación de Investigadores COLCIENCIAS 2012" and "Convocatoria de apoyo a la movilidad de estudiantes de la FIA 2013-2014 de la Universidad Nacional de Colombia - Sede Manizales".

References

1. Trifa, V.M., Girod, L., Collier, T., Blumstein, D.T., Taylor, C.E.: Automated wildlife monitoring using self-configuring sensor networks deployed in natural habitats. In: International Symposium on Artificial Life and Robotics (AROB 2007), Beppu, Japan (2007)
2. Harma, A.: Automatic identification of bird species based on sinusoidal modeling of syllables. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE (2003)
3. Neal, L., Briggs, F., Raich, R., Fern, X.Z.: Time-frequency segmentation of bird song in noisy acoustic environments. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2012–2015 (2011)
4. Fagerlund, S.: Bird species recognition using support vector machines. EURASIP Journal on Advances in Signal Processing 2007 (2007)
5. Trifa, V.M., Kirschel, A.N.G., Taylor, C.E., Vallejo, E.E.: Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. The Journal of the Acoustical Society of America 123(4), 2424–2431 (2008)
6. Briggs, F., Fern, X., Raich, R.: Acoustic classification of bird species from syllables: an empirical study. Technical report, Oregon State University (2009)
7. Stowell, D., Plumbley, M.D.: Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers. Technical report (2011)
8. Chu, W., Blumstein, D.T.: Noise robust bird song detection using syllable pattern-based hidden Markov models. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 345–348 (2011)
9. Hao, Y., Campana, B., Keogh, E.: Monitoring and Mining Animal Sounds in Visual Space. Journal of Insect Behavior 26(4), 466–493 (2012)
10. Huang, C.J., Yang, Y.J., Yang, D.X., Chen, Y.J.: Frog classification using machine learning techniques. Expert Systems with Applications 36(2), 3737–3743 (2009)