DETECTION OF BIRD VOCALIZATION ON FIELD RECORDINGS: A COMPARISON OF METHODS

Juan Sebastian Ulloa, Hernán Darío Benítez Restrepo*

Department of Electronics and Computer Science, Pontificia Universidad Javeriana.
Calle 18 No. 118-250 Cali, Colombia. *Electronic mail: jsulloa@javerianacali.edu.co

ABSTRACT

Three signal detection algorithms were adapted to find bird vocalizations on field recordings: an energy threshold, a neural network (9 acoustic features) and a template matching approach. From a sound library, 433 audio samples (335 bird vocalizations and 98 interference sounds) from the Colombian Paramo and high-Andean forest were extracted. Two different tests were designed, the first one was focused on estimating the sensitivity of the algorithms, and the second one to analyse their specificity. The energy threshold showed to be a good starting point for analysing the signals when interference is scarce. The template matching is a robust alternative and is a method to be used in stereotyped birds' vocal sounds. Finally, the neural network is effective to discriminate between vocalizations and interferences, but sensitive to ambient noise.

Index Terms — neural networks, energy threshold, template matching, sensitivity, specificity.

1. INTRODUCTION

Birds are valuable indicators of biodiversity and environmental health of terrestrial habitats [7]. Their vocalizations are a source of information that can be used to explore the composition and diversity of birds in a particular area of interest [3], notably in tropical countries, where dense foliage limits visibility [20], [14]. New technologies have allowed automated recording systems that perform continuous long-term recordings. However, the analysis of such recordings is time consuming and requires significant human effort. Using signal processing and pattern recognition techniques a system can identify and discriminate sounds, reducing time and scientific research costs.

The signal recognition process is divided into two successive tasks: *detection* and *classification* of data. The aim of signal detection is to identify and separate structured sounds of interest from noisy background. On the other hand, classification seeks to subdivide the signals detected in biologically relevant groups. Birdsong recognition relies on segmentation as an early processing step. This can be done manually, but with large datasets proper automatic recognition should be done automatically. An imprecise detection generates noise in the whole system and affects the recognition results [8]. Hence, the good performance of the detection algorithm is an essential step for subsequent parts of the study. The present work focuses on the detection step.

The simplest and most common algorithm to detect a signal against background noise in acoustic recordings is the threshold energy. This technique has been used in speech processing to locate the beginning and end of voice sections [15], [9]. In animal bioacoustics, Fagerlund [8], Somervuo *et al* [17] and Briggs *et al* [4] used the energy threshold approach to identify and segment bird vocalizations. Its execution speed, allowed to apply this technique in real-time applications monitoring birds [19] and marmots [1]. Finally, it should be added that this technique has now been incorporated into commercial software that is used by the scientific community: *Avisoft Bioacoustics* (http://www.avisoft.com/).

Advanced techniques involve more sophisticated classifiers (detection can be regarded as a binary classifier) that combine a greater number of features. Skowronski and Harris used Gaussian mixture models to detect and classify echolocation calls of bats [16]. Neal *et al* applied a Random Forest algorithm –multiple decision trees– for visual analysis of spectrograms to identify bird vocalizations [13].

Another relevant approach is to use the spectrogram cross-correlation. This technique has been implemented to detect whales on underwater recordings from the North Pacific Ocean [11], [12].

Just a few studies, [16], [11], have undertaken a comparison of methods to evaluate detection performance. Ground truth was established labelling individual calls by hand. For the present study, a new approach to evaluate the algorithms was designed. Specificity and sensibility tests were implemented in order to assess the performance of the algorithms. The aim of this paper is to shows the advantages and disadvantages of three algorithms adapted to detect bird vocalizations on field recordings collected at the Paramo and high-Andean forest: an energy threshold detector, a neural network and a template matching approach.

Sample Type	Number of samples	Sound Type
VOCALIZATIONS		
Acropternis orthonyx	58	Quasi-constant frequency
Atlapetes pallidinucha	57	Frequency modulated
Atlapetes schistaceus	46	Frequency modulated
Atlapetes torquatus	44	Frequency modulated
Cinnycerthia unirufa	41	Broadband pulse
Henicorhina leucophrys	52	Frequency modulated
Penelope montagnii	37	Broadband pulse
Total	**335**	
INTERFERENCES		
Clicks and pops	54	Broadband pulse
Human Voice	44	Rich harmonic content
Total	**98**	
TOTAL SAMPLES	**433**	

Table 1: Details of sounds selected for this work: vocalizations of seven species of birds and two types of interference.

2. MATERIALS AND METHODS

2.1. Database

To develop the algorithms and create a testing protocol, a sound database was consolidated. The *Instituto de Recursos Biológicos Alexander von Humboldt* provided the sound recordings from their *Colección de Sonidos Ambientales*. First, recordings from the Paramo and the high-Andean forest were chosen. Then, only bird vocalization with no overlap and with more than 25dB of signal-to-noise ratio were selected. Three broad categories of bird sounds were included on the database: quasi-constant frequency, frequency modulated whistle, and broadband pulses.

Furthermore, possible interferences were identified. At high altitude, over 2800m above see level, birds dominate the day ambience sound. Diurnal, insects and amphibians sounds, are scarce. Interferences come from the recording equipment handling –two common types, `clicks' and `pops'– and from the human voice announcing complementary data during the recording. Table 1 shows the details of the final database.

2.2. Constant False Alarm Rate (CFAR)

The CFAR is an adaptive energy threshold algorithm. It was implemented to detect bird vocalizations by Vlad Trifa [19]. The program identifies high-energy segments on audio signals. First, it estimates the statistical distribution of the amount of energy during N consecutive samples. Ambient noise is assumed to have a normal distribution $N(\mu,\sigma)$. With this estimate, a threshold is set at β standard deviations over the mean energy $(th= \mu + \beta\sigma)$.

The audio signal was discretized into 20ms bins and in each bin the energy level is calculated. If the signal energy exceeds the threshold previously defined, the section is labelled as a bird vocalization. An additional discriminative feature was added, minimal duration of the segment (*l_voc=40ms*). This feature helps to make distinctions between `click' and `pop' interferences, and bird vocalizations. Thus, an event needs to have a high-energy segment and a minimum length of 40ms to be labelled as a bird vocalization.

2.3. Artificial neural networks (ANN)

Artificial neural networks are computer models inspired on biologic neural networks. To face the problem of bird vocalization detection, a general ANN model was implemented, the *feed-forward perceptron multilayer* network [6], as a binary classifier. The net has three levels: an input layer, then 40 neurons on the hidden layer and finally an output layer. A sigmoidal function activates the neurons on the hidden layer, while the output neuron has a linear function. Nine acoustic features were used on the model: seven spectral features (spectral centroid, spectral roll off, spectral flux, spectral entropy, frequency bandwidth, peak frequency, spectral flatness) and 2 temporal features (short time energy, zero crossing rate). More information about these features can be found on [8], [9] and [18].

The *Backpropagation* algorithm was used to train the ANN. A balanced training set included: (1) 257 samples of bird vocalizations representing 54 seconds of audio and (2) a mix of pink noise and interferences of 57 seconds length. Noise samples were divided into 40 seconds of pink noise, 41 recording equipment handling interferences and 35 segments of human voice. All the samples were added to form an audio track and the acoustic features were calculated on frames of 20 milliseconds. Thus, each object of the neural network is a 20ms audio frame defined by the calculated features.

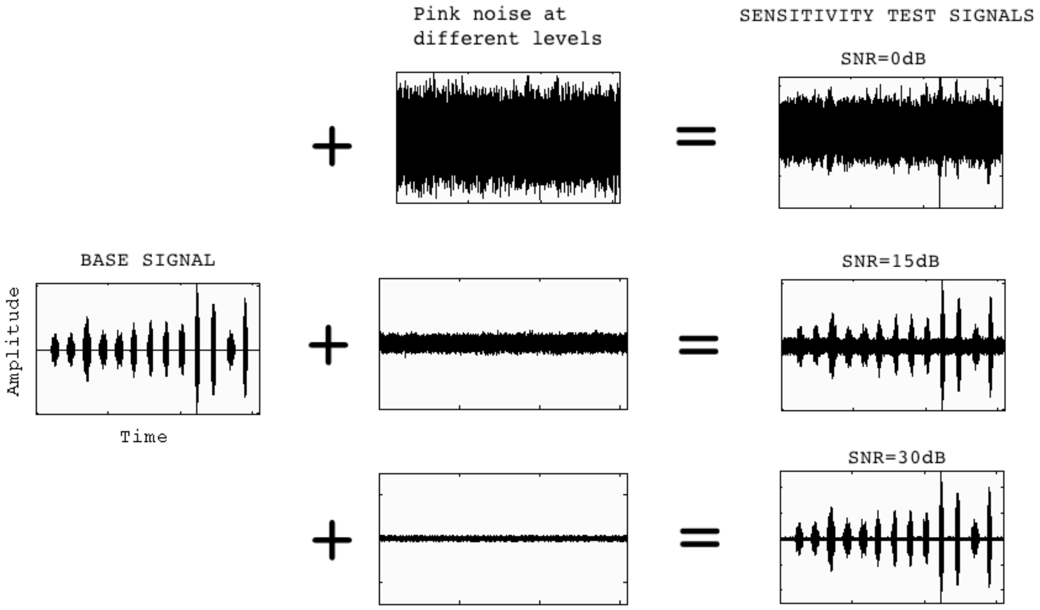


Figure 1: Schematic of sensitivity test signal synthesis

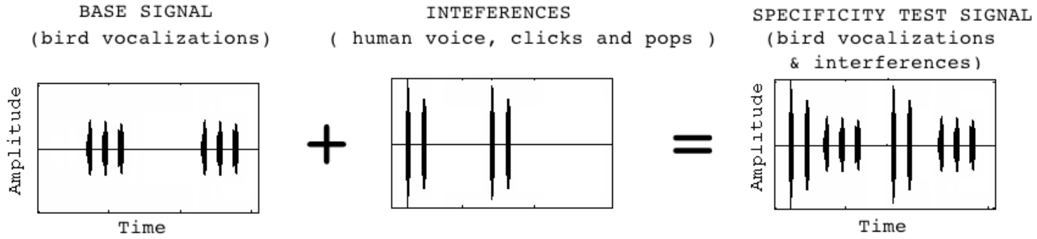


Figure 2: Schematic of specificity test signal synthesis

2.4. Template Matching (TM)

Template matching is used to find areas of an image that are similar to a template [5]. First a spectrogram from the audio signal was generated. Searching for a balance between temporal resolution, spectral resolution and bandwidth the following parameters were used: 512 point FFT, 512 frame size, Hamming window and 75% overlap between frames. This technique was used for stereotyped birds' vocal sounds, more precisely focused on finding *Acropternis orthonyx* vocalizations. The template was created with one sample with good signal-to-noise ratio (30dB).

The cross-correlation signal is a measure of similarity between the template and the field recording at a precise location. Hence, a peak will occur in the output signal for every target signal that is present in the recording. To segment the audio into signals of interest, a threshold at 0.5 on the output signal was defined and peaks over that threshold were selected. The length of the template signal gave the length of the segment. Thus, the recognition function became a sequence of discrete detection events.

3. TESTS AND RESULTS

3.1. Testing model

Algorithms were evaluated with systematic sensibility and specificity tests. In order to control different variables and to establish a ground truth, artificial test signals were created. On one hand, bird vocalizations and interferences were segmented from field recordings. On the other hand, the ambient noise was synthesized with computational methods: creating white noise and applying a -6dB per octave low pass filter, the process gives a pink noise signal. Test signal result from mixing bird vocalizations, interferences and pink noise in different ways.

The performance of a detector is measured through two statistical variables: sensitivity and specificity. Hence, tow different tests were performed. The first one is focused on assessing the sensitivity of the algorithms; it comprises bird vocalizations and variable ambient noise. For this, three signals were made with different signal-to-noise ratio: 30dB, 15dB and 0dB. The second test examines the specificity. It has bird vocalizations and interferences ('clicks', 'pops' and human voice). Slight pink noise is added (SNR=30dB) to simulate the inevitable ambient noise in audio recordings, but this time the signal-to-noise ratio remains constant. Figures 1 and 2 illustrate this process.

In addition, a second round of tests was performed with the addition of a pre-processing step. As the bird vocalization on the database had a minimum frequency of 1.2 kHz and a maximum frequency of 9.3 kHz, a digital FIR filter was implemented with pass-band frequency between 1 and 10kHz.

3.2. Results

The Receiver Operating Characteristic (ROC) curve for each detector was created defining minimum and maximum thresholds –zero false negatives and zero false positives– and then computing the output of 100 linearly spaced thresholds between the extremes. To measure quantitatively the performance of the detectors, time segments were discretized in 20 milliseconds bins. The detectors determine if at least a fraction of a vocalization occurs in each bin and the results are compared with the ideal output given by vocalization segments in the test signals. The Area Under the Curve (AUC) gives the performance of the algorithms overall accuracy [2].

Sensitivity test: Figure 3 shows ROC curves for the detectors and a summary of the AUC results in a bar chart. The performance of the three algorithms is similar when the signal-to-noise ratio is high, however, clear differences come into view as the ratio diminishes. The performance of the template matching technique, between SNR=30dB and SNR=0dB, decreases only 2.8%, while the energy threshold showed 31.5% and the neural network 49.5%. Therefore, it appears that the ambient noise affects more the neural network and the energy threshold than the template matching algorithm.

Specificity test: The results of this test, showed that the neural network and the template matching approach had little variation on their performance, decrease of 2.5% and 3.4% respectively. Instead, the performance of the energy threshold approach fell 58.6% (Figure 4).

Addition of the band-pass filter: With the inclusion of a pre-processing step, the sensibility and specificity tests were repeated. On average, the neural networks improved by 11.1% and the energy threshold 20.0%. In contrast, the performance of the template matching had little variation, it only increased 1.03%. The trend on both tests remains the same. On the sensitivity test, the template matching had the lowest variation on it's performance with 2.1%, followed by the energy threshold with 6.0% and the neural network was the most affected by the background noise with 16.0%. About specificity, the best algorithms were the template matching, with a particular increase of 0.8%, and the neural network with a decrease of only 1.7%. Finally, the energy threshold performance dropped by 4.9%. The results are shown on Figures 5 and 6.

3.3. Execution time

A detailed run-time analysis of the algorithms is beyond the scope of this research. Some execution times of the different techniques are presented to give an idea of the order of magnitude of these values. The field recording *Acropternis_orthonyx_10880.wav* (95.0665 seconds length) was analysed by each algorithm in a 2.4GHz processor and 8GB memory computer. The codes were implemented in the MATLAB® R2011a software (The MathWorks, http://www.mathworks.com). Results are presented in Table 2.

	CFAR	RNA	TM
Execution time (s)	0.503	13.508	3.552

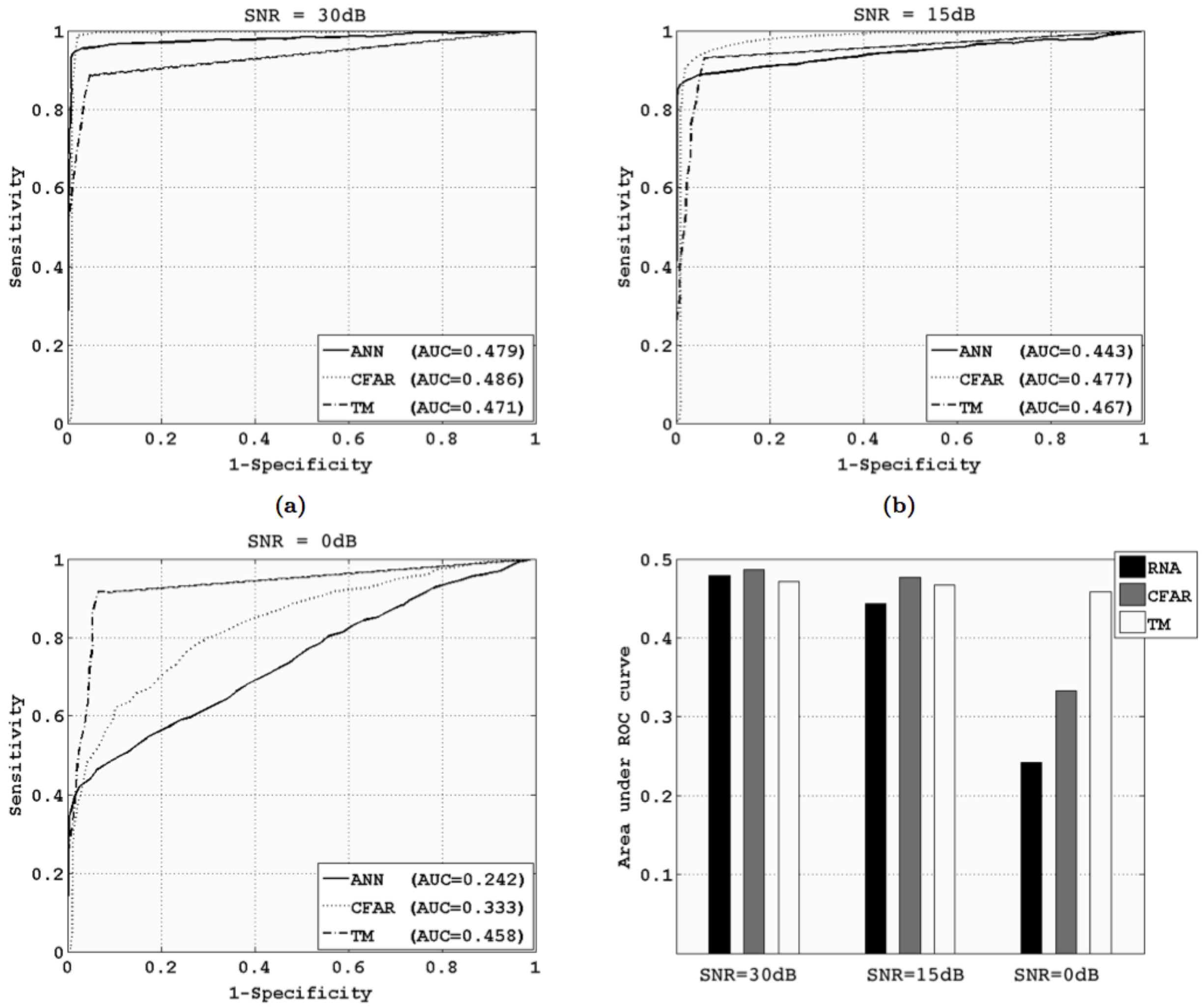
Table 2: Execution time of each algorithm analysing a 95s field recording: Acropternis_orthonyx_10880.wav.

4. DISCUSSION AND CONCLUSIONS

This work contributes to the comparison of different techniques to detect bird vocalizations. It presents a testing model that quantitatively reveals advantages and disadvantages of the techniques. In a similar way that Skowronski and Harris did for bat calls [16], we discretized time in 20ms frames to quantify the performance. However, we used artificial signals to make a more exhaustive and precise test of the algorithms. In [16], ground truth was established hand labelling the beginning and end of the signal. For bird vocalizations, we found that it was ambiguous to precisely define these boundaries, hence the signal was segmented from the original recording were the vocalization was clearly distinguishable. As the testing is composed of these segments, we could have definite ground truth. The advantages and disadvantages of the algorithms are highlighted in this section. Table 3 shows a summary of the analysis.

Threshold energy Although it is a simple technique, this algorithm is optimal for signals with high signal-to-noise ratio. The effectiveness is affected by ambient noise and more critically when facing interferences. This can be inferred by the working mechanism, any high-energy signal will be detected regardless if it's a bird vocalization or an interference, increasing the false positive rate. A band-pass filter can make the algorithm more robust. The interferences used in this work, and most of the anthropogenic noise, have their energy in low frequencies. Therefore, the filter can attenuate them. It should be noted that in cases where interference can't be attenuated by a filter, the false positive rate will increase as this is not a specific detector.

Artificial Neural Network. This technique is very effective with signals that have interferences, without the need of an additional filter (without filter AUC = 0.467,



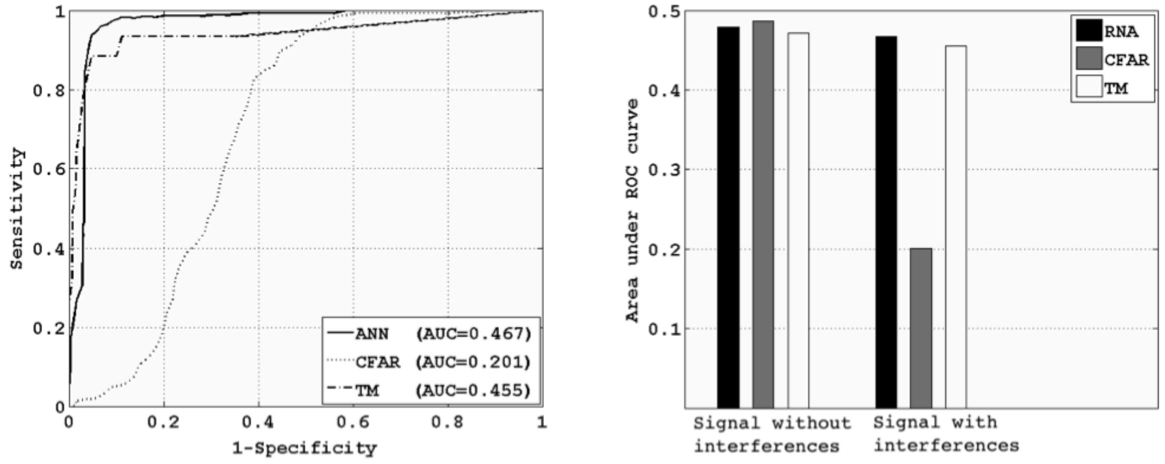
(a)

(b)

(c)

(d)

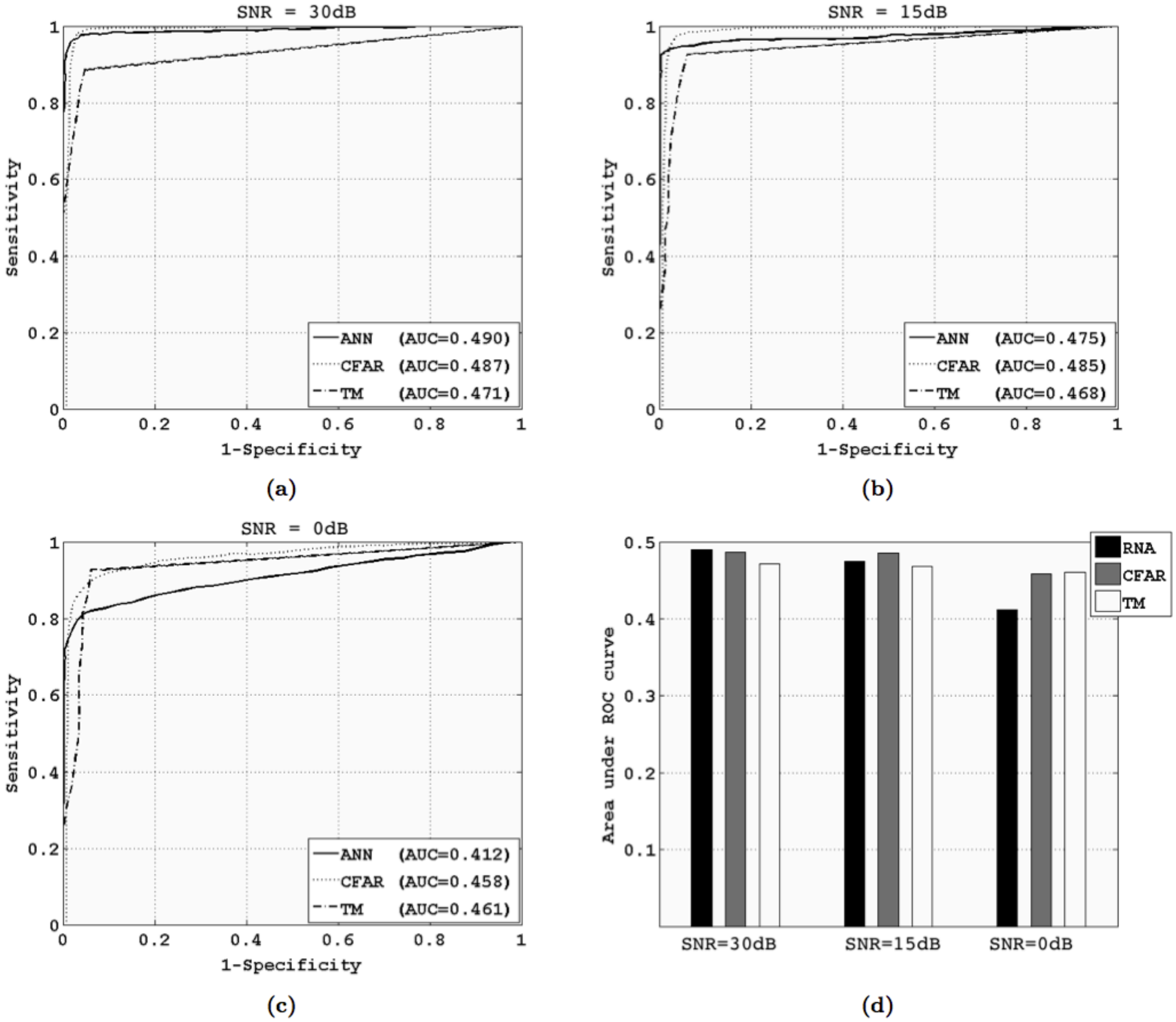
Figure 3: Sensitivity test. ROC curve results (a, b y c) and AUC bar chart (d).



(a)

(b)

Figure 4: Specificity test. ROC curve results (a, b y c) and AUC bar chart (d).



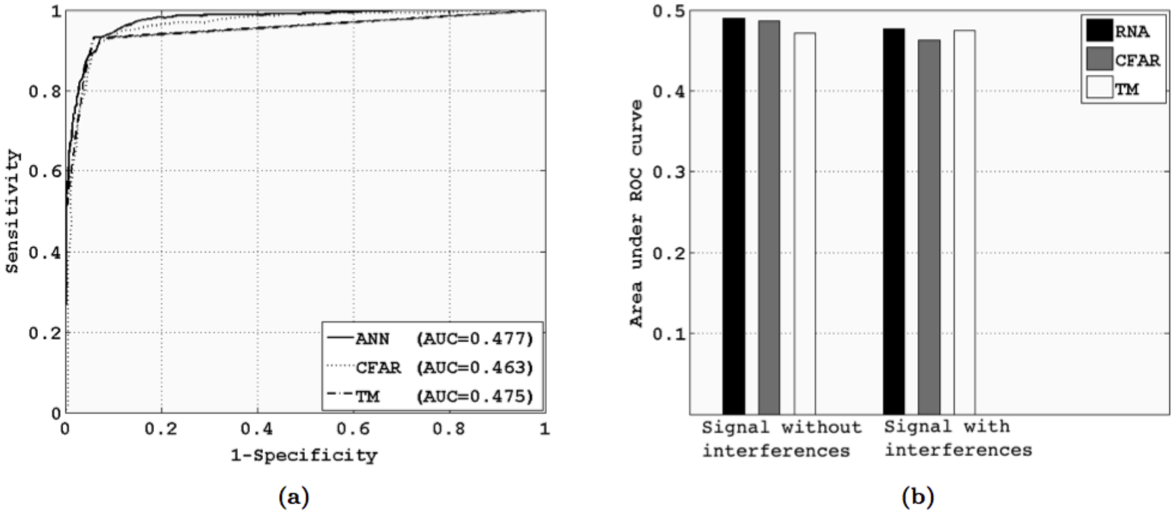
(a)

(b)

(c)

(d)

Figure 5: Sensitivity test with band-pass filter. ROC curve results (a, b y c) and AUC bar chart (d).



(a)

(b)

Figure 6: Specificity test with band-pass filter. ROC curve results (a, b y c) and AUC bar chart (d).

with filter AUC = 0.477). However, the performance is affected severely with the presence of ambient noise. The results showed that the neural network was more affected by ambient noise than the energy threshold and the template matching approach. On the contrary, Skowronski and Harris showed that machine learning algorithms had a better performance for faint signals [16], and Mellinger and Clark had better results with neural networks than spectrogram correlation [11]. Our lower performance might be caused by acoustic features used. These, capture the information of the entire spectrum that may be mainly dominated by noise since bird vocalizations are often narrow band signals. This also explains why in other studies, Mel Frequency Cepstral Coefficient features do not work effectively to distinguish bird sounds, as observed by Jančovič and Köküer [10]. This hypothesis need a deeper study to be confirmed; the functioning of a trained network is difficult to understand, making it uncertain under what conditions the network might fail. Additionally, it's important to note that ANN, and machine learning algorithms in general, requires large training data set and associated operator time for data preparation.

Template matching. The algorithm shows good results, although not outstanding. It's maximum AUC is 0.475 and all the ROC curves show that a 100% of sensibility is only achieved with high false positives. This is a specialized detector for a particular vocalization and is not adapted to changes in time or frequency. However, the minimum AUC value on all tests was 0.455. This shows that this detector is robust to ambient noise and interferences from field recordings. Therefore, this is a suitable technique to be implemented on stereotyped bird vocalizations.

	CFAR	RNA	TM
Robust to ambient noise	★★★	★★	★★★★
Robust to interferences	★	★★★★★	★★★★
Easy to implement	★★★★★	★	★★★
Execution speed	★★★★★	★★	★★★
Adaptability	★★★★★	★★★★	★

Table 3: Five characteristics are rated on a scale from one star (⋆) to five stars (⋆⋆⋆⋆⋆), five stars being the highest score.

By adding the band-pass filter, the results of the energy threshold and the neural network improved substantially. The use of frequency filters are encouraged since they are easy to implement and they allow to focus the analysis of signal processing algorithms.

On manual field recordings, signals of interest are recorded with high signal-to-noise ratio and interferences produced by the recording equipment are common. Hence, The neural network is an interesting option to segment the recordings for posterior classification, although it demands more computational effort than the other alternatives. On the other hand, long-term acoustic monitoring programs, microphones collect ambient sound. The signal-to-noise ratio will vary continuously, depending on the weather (rain, wind) and the distance between the sound source (the bird) and the microphone. The Paramo and high-Andean environments have soundscapes where the bird vocalizations have no significant interferences. Hence, the energy threshold plus a band-pass filter are a good starting point to analyse sounds recorded on these ecosystems. The template matching algorithm is adapted for particular sounds, although this limits its application it can be used to search for rare and cryptic species with stereotyped vocalizations (either the song or only part of it). Neural network –with the acoustic features used here– showed to be very sensible to ambient noise and is the least recommendable to analyse environmental acoustic data with little interferences.

For execution time and improved detection purposes, a combination of algorithms could be implemented. For example, as a first step, identify signals with high energy using the energy threshold and then, as a second step, analysing those segments with the neural network to discriminate between interferences and bird vocalizations, reducing the false positive rate.

Applying signal processing algorithms will allow to save time and effort of trained personnel. With tight budgets and overwhelming biodiversity, engineering must bring more tools to analyse environmental acoustic data in tropical ecosystems.

ACKNOWLEDGEMENTS

This work was supported by the Colombia Biodiversa grants from the Alejandro Angel Escobar Foundation. Audio recordings used in this project were taken from the Colección de Sonidos Ambientales from the Instituto Alexander von Humboldt Colombia. The authors would like to thank Paula Caicedo, Jaime Parra and Mauricio Orozco for their assistance on this research.

REFERENCES

[1] A.M. Ali, K. Yao, Travis C. Collier, C.E. Taylor, D.T. Blumstein, and L. Girod. “An empirical study of collaborative acoustic source localization,” *In Proceedings of the 6th international conference on Information processing in sensor networks,* pages 41–50, 2007.

[2] A.P. Bradey. “The use of the area under the roc curve in the evaluation of machine learning algorithms,” *Pattern Recognition*, 30(7):1145–1159, 1997.

[3] T. S. Brandes. “Automated sound recording and analysis techniques for bird surveys and conservation,” *Bird Conservation International*, 18:163–173, 2008.

[4] F. Briggs, Raviv Raich, and Xiaoli Z. Fern. “Audio classification of bird species: a statistical manifold

approach," *In Data Mining, Ninth IEEE International Conference*, 2009.

[5] R. Brunelli. *Template Matching Techniques in Computer Vision: Theory and Practice*. John Wiley and Sons, Ltd, 2009.

[6] E.F. Caicedo and J.A. López. *Una aproximación práctica a las Redes Neuronales Artificiales.* Programa Editorial Universidad del Valle, 2009.

[7] V. Carignan and M Villard. "Selecting indicator species to monitor ecological integrity: a review," *Environmental Monitoring and Assessment*, 78:45–61, 2002.

[8] S. Fagerlund. "Automatic recognition of bird species by their sounds," *Master's thesis, Helsinky University of Technology,* Nov. 2004.

[9] T.D. Giannakopoulos. "Study and application of acoustic information for the detection of harmful content, and fusion with visual information," *PhD thesis, National and Kapodistrian University of Athens*, 2009.

[10] P. Jančovič and M. Köküer. "Automatic detection and recognition of tonal bird sounds in noisy environments," *EURASIP Journal on Advances in Signal Processing*, 2011(982936), 2011.

[11] D. K. Mellinger and C.W. Clark. "Recognizing transient low-frequency whale sounds by spectrogram correlation," *Journal of the Acoustical Society of America*, 107(6):3518–3528, June 2000.

[12] L.M. Munger, D.K. Mellingerand S.M. Wiggins, S.E. Moore, and J.A. Hildebrand. "Performance of spectrogram cross-correlation in detecting right whale calls in long-term recordings from the bering sea," *Acoustique canadienne*, 33(2):22–27, 2005.

[13] L. Neal, F. Briggs, R. Raich, and X.Z. Fern. "Time-frequency segmentation of bird song in noisy acoustic environments," *In Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference*, pages 2012–2015, 2011.

[14] T.A. Parker. "On the use of tape recorders in avifaunal surveys," *Auk*, 108:443–444, 1991.

[15] L.R. Rabiner and M.R. Sambur. "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, 54:297–315, Feb. 1975.

[16] M.D. Skowronski and J.G. Harris. "Acoustic detection and classification of microchiroptera using machine learning: Lessons learned from automatic speech recognition," *Journal of the Acoustical Society of America,* 119(3):1817–1833, 2006.

[17] P. Somervuo, A. Härmä, and S. Fagerlund. "Parametric representations of bird sounds for automatic species recognition," *IEEE Transactions on audio, speech, and language processing*, 14(6):2252–2263, 2006.

[18] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, third edition, 2006.

[19] V.M. Trifa. "A framework for bird songs detection, recognition and localization using acoustic sensor networks," *Master's thesis, École Polytechnique Fédérale de Lausanne*, Feb 2006.

[20] H. Villarreal, M. Álvarez, S. Córdoba, F. Escobar, G. Fagua, F. Gast, H. Mendoza, M. Ospina, and A.M. Umaña. *Manual de médodos para el desarrollo de inventarios de biodiversidad. programa de inventarios de biodiversidad.* Instituto de Investigación de Recursos Biológicos Alexander von Humboldt, Segunda edición:236, 2006.