2015 4th International Conference on Modeling and Simulation

Bird Sounds Detection Using Normalized Audio Power Sequences

Babacar Diop[+], Dame Diongue[*], Ousmane Thiare[+]

[+]Faculty of Applied Sciences and Technology, UGB, BP. 234 Saint-Louis, Senegal

Email: {diop.babacar, ousmane.thiare}@ugb.edu.sn

[*]Faculty of Applied Sciences and Technology, UASZ, BP. 523 Ziguinchor, Senegal

Email: dame.diongue@univ-zig.sn

Abstract

In this paper, we bring out the first research element in building an automated surveillance system for rice field monitoring against pest and grain-eating birds. We study how accurate can be the use of normalized power sequences in detecting bird sounds, which is a kind of harmonic sound type. We proposed a simply detection scheme (which can be executed by sensors), for the purpose of detecting the presence of birds based on the sounds and calls they produce. The scheme is based on pruning audio frame blocks to keep relevant peaks, computing the normalized power of the sequences of captured audio files, and derive the presence of birds by only considering important values of the variances of the input files. The experimental results conducted to test the efficiency of the scheme show that 91,07% of bird calls from our database can be correctly identified.

Keywords: Wireless sensor networks - Bird sounds detection - Multimedia surveillance - Normalized power signal.

I. Introduction

The use of acoustic sensors is a very efficient way to monitor several types of events, provided that audio data are generated into the monitored space by specified sources. A generic system consists of exploiting audio data captured by acoustic sensors from target sources, signaling events occurrence to an automated system in cases of critical events, and then performing further processing for concluding on the presence of a target, and for localizing the acoustic source [1]–[4].

In this paper we propose a detection system for the purpose of detecting bird calls, by training a data set of audio files. This study stands as the first research element in conceiving an automatic detection system for efficiently monitoring rice fields against the intrusion of grain-eating birds. Our data set has been constituted in two steps. We first recorded more than five hours of audio sounds from a village, close to *Djoudj*'s International Protected Bird sanctuary, which includes a huge number of grain-eating birds - the presence of water in the surrounding combined to the natural appeal of the environment encourage the establishment of several birds' species. Then, in the second step, we synthesized more than five hours of bird sounds by combining components that were downloaded from Internet, and generated noise signal with different means, variances and power levels. For the detection system, we first consider frame blocks of approximately 900 milliseconds each, that are cleansed by considering a threshold signal value modeled from a reduction factor r, and then applied a detection system based on the computation of the normalized power sequences on each sub-frame of 20 milliseconds and the variance [4]. A threshold value is assigned to the variance to detect or not occurrences of bird calls in the tested frame block.

The rest of the paper is organized as follow - Section II provides a slight overview about bird vocalization system to highlight some informative properties related to our detection scheme. In section III we give detailed explanation of the proposed detection scheme which is divided into two phases, and describe the different stages included in each phase. Results obtained from the

978-1-4673-9828-2/15 $31.00 © 2015 IEEE
DOI 10.1109/MAS.2015.20

experiments held upon the training data set are presented and interpreted in section IV. Finally section V concludes the paper with some perspectives.

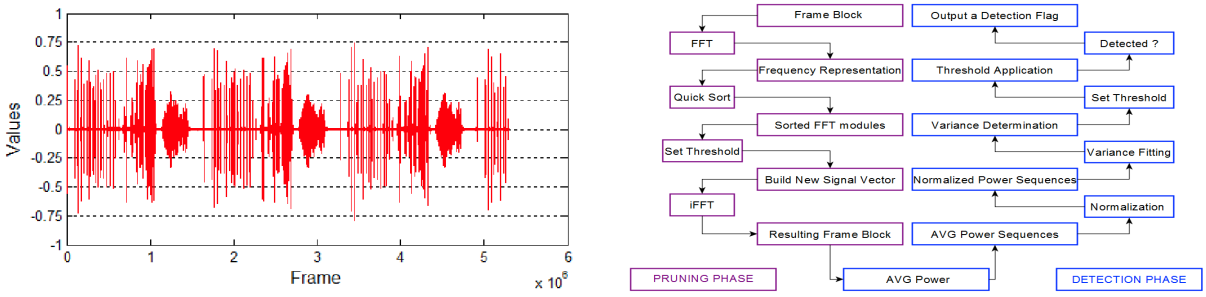
II. Bird Calls and Songs

Bird sounds can be categorized into two classes namely - calls and songs [5]. Both classes of sounds may be produced for special purposes (e.g. communication, call to copulate).

Bird songs are usually characterized as being complex and long. Considering the sex of the sound producer and the purpose of the produced sound, bird song may be more or less complex (and long). For instances the sounds produced by birds for female attraction are longer and less complex than the song for a call for mobilization; female songs are simpler than male's one in different species. Birds use to sing at dawn for it is the best time to copulate with females, to take vacant territories and for the early day feeding conditions.

Contrasting with bird songs, bird calls are most of time short and simply. According to [4] they can sometimes be complex or confused. A very important property is that putting aside the gender aspect, bird calls typically occur in contextual cases, with some specific functionality. Among the incomplete possible set of functionality, we target the feeding and alarm calls for the purpose of controlling their intrusion within a feeding space (case of a rice field).

Since bird calls are most of times less complex and shorter than bird songs, one can detect the presence of a bird call by identifying temporal pieces of notes, which can be referred as syllabic elements, that are iterated during the sound production to form the whole bird call. By exploring bird sounds database, with more above 75% especially comprised of *Quelea Quelea* and *Passer*'s sounds, the call duration, materialized by one short whistling, ranges between 600 and 900 milliseconds. Lying under this range, it is possible that one second might be sufficient for detecting a single bird call.



(a) Variation of a single bird call followed by several bird calls as a superposition of observed peaks with the single bird song.

(b) Flowchart showing different steps of the detection feature process.

Fig. 1. Bird Calls Performing and Feature Processing Description.

Figure 1.(a) presents a bird call instance, recorded from a rice field, where one single bird starts the appeal, and then followed by other birds. We can see that during the period where different birds are whistling together, the plotted dark lines result from the superposition of individual calls performed by single birds. The temporal pieces of notes that we refer to above, can be identified in figure 1 by considering the three sequences respectively from 1 to 1.5×10^6, 1.6×10^6 to 3.1×10^6 and 3.3×10^6 to 4.7×10^6. Each piece of note is comprised of approximately 1.4 to 1.5×10^6 audio elements, corresponding approximately to 36 seconds. By plotting the spectrum of our bird calls data set, we found several pieces of calls corresponding to different values between 3 seconds to more than 2 minutes.

III. Detection of Bird Sounds

In this section we present the system proposed in this paper to detect bird calls. The detection system consists of two phases as described in Figure 1.(b) - a pruning phase during which non valuable values are removed from the frame block, and a detection phase where pieces of bird sounds can be identified in the frame block by computing and normalizing the average power sequences, and using variance handling technique to highlight relevant peaks. Below we give details about the different stages of the detection system for these two phases.

A. Pruning Phase

The idea behind this first phase is to initially proceed in a data cleansing stage. This goes by converting digital signal values into frequency representation, and pruning unattractive values - i.e the values whose presence induces no known benefit for the performed operation - and to only retain relevant values. All relevant values must fluctuate above a threshold value S^{th} that depends on a predetermined reduction factor r. This yields a reduction of the observed frame block and make the detection process more accurate and more convenient. We propose to perform the pruning phase in five steps -

1) Transform numerical data into frequency components by using FFT algorithm
2) Sort modules of the Fourier transform with the quick sort algorithm
3) Compute a threshold signal S^{th} according to the reduction parameter r
4) Build a new signal vector, keeping only signal values above S^{th}
5) Reconstitute the signal into digital values by using inverse FFT algorithm

The first and the fifth steps are realized using respectively the FFT (Fast Fourier Transform) and the inverse FFT algorithms. The threshold is computed on the basis of formula (1), where L is the length of the frame block and r as mentioned above is the reduction factor.

$$S^{th} = L - L/r \tag{1}$$

A very important stage is the determination of the reduction factor r that will fit the recorded data. We proceed by iterating the reduction parameter from 2 to 10, and observed the resulting signal visually and by listening to it. The conventional values of the reduction factor that we kept from this process range between 3 and 5. However this value may deviate considering the level of the signal. That is with signal-to-noise ratio less than 0, the value of the reduction factor may increase. The general remark is that greater the noise level, then the greater the reduction factor must be. Finally the choice of the quick-sort algorithm is to gain in run time complexity

B. Detection Phase

The detection system that we present here is based on the computation of the power sequences for each sub-frame block of 2,000 elements. Then the obtained values are normalized to calibrate the range between 0 and 1, and the variance is computed for each frame block comprised of 20 sub-frames. Finally a threshold is applied on the variance to conclude on the positiveness of the detection of bird call instances.

1) Signal power calculation : Signal power and energy are widely used characteristics in signal processing. The power of an audio signal is usually computed as the average power within a period without taking in account the total time duration. In our proposed detection scheme, we compute the power signal, denoted as $P(X)$ as the sum-of-squares of discrete values, ranging in a frame of N samples. Then the power of an audio signal is the sum of the absolute squares of its time-domain samples divided by the signal length. Summing absolute squared values would prevent the power of a symmetrical waveform to take the value zero.

According to the specificity of the targeted application, we set the length N of a frame block to 2,000 samples, corresponding to 45 ms. So the signal power is estimated for each frame block.

2) Normalization of power sequences: The second step of the detection scheme consists of computing normalized values of each power block f, within a chosen window. Fortunately thinking, the use of windowing technique is not a requirement of our design scheme. We do not need to take the most recent values or to reshape the window's dimension upon which each normalized value is computed. This is due to the fact that the length of each frame block within which an average power is obtained, is respectful with regard to the frame length that can be captured and processed by an acoustic sensor in wireless sensor network based surveillance system. The normalized power value for a frame block is given in [4].

3) Finding a threshold: At this step, the variance is computed for the L values of the captured frame. The use of the variance, derived from the power sequences normalization is one way of detecting a sudden power change. When computing the variance of the captured frame, the last power value is not included in the variance estimation. By doing so, the normalization step which leads to a recalibration of computed average power values between 0 and 1, with a slight variance evolution, transforms this latter into a very powerful detection technique, since a sudden change can be induced by a sudden deviation of the variance.

The smaller the length of a frame block, the more details will appear for the calculation of the variance. So in this case, complexity is greater and sometimes so much details may prevent to consider useful values and then could lead to confuse results. In the opposite, the greater the length of a frame, the shorter the run time complexity. With large frame length, computed variances may be less accurate.

Normally the length of a frame should be a measure relative to the temporal properties of the type of sound to detect. For example, considering impulsive noise (like dog barking) and tonal noise (bird calling) detections, the length of each one is proportionate to the fraction of time that is required to identify the sound nature, and to statistically deduce the induced change considering the whole frame block.

Two cases are relevant in our context considering what is stated above.

- The first is that when detecting bird calls in the presence of silence and even in slightly noisy conditions, we expect to get an acceptable sensibility of the variance evolution based detection mechanism, even if several birds are simultaneously performing calls. But we can denote an increase of the intensity of the mitigated sounds, as described in Figure 1, which positively influences the variation of the variance. In this case, we can expect that whenever the variance of the current frame block is greater than the minimum variance level in silence and slightly noisy conditions, then a positive detection flag can be output.
- The second is linked to the difficulties in detecting sounds in very noisy environments where sounds of different frequencies may colorize the tone of bird calls. This guides to very confuse results since one can not lean upon the minimum variance driven in silence and slightly noisy conditions to detect change in the variance spectrum. In this case, the system must keep track of the recent values of variances from the initial silence and slightly noisy conditions to activate positive detection flags, each time an obtained variance is above those values.

IV. Experimental Results

The performance of the proposed scheme is tested upon a database containing recorded audio elements (cf. Introduction).

We carry the simulations with two objectives - the first one is to evaluate the accuracy of the proposed scheme in detecting bird calls considering high SNR ratio, the second is to study its ability to provide acceptable results in harsh conditions, with reducing SNR ratio. We synthesized specific audio files for the purpose of simulating harsh environments recordings, where the intensity of the sound signal (bird call) may be highly reduced. For this, we generate Gaussian noise by specifying the mean, variance and power for each tested frame.

A. Audio Segmentation

We begin to segment the recorded audio files into frame blocks of 40,000 elements each, corresponding to 900 ms. Then the pruning phase is applied. As mentioned above, the application of the pruning phase refines the way the audio feature extraction is made up. We can remark that the pruning phase is executed just once considering the whole frame length (of 40,000 elements), because of the run time complexity it will induce while being applied for each sub frame of 2,000 elements. We used *fft* and *ifft* algorithms from Matlab, and implemented the quick-sort algorithm.

In the second phase, each frame block of 40,000 length is split into 20 sub-frames of 45 ms. This corresponds to approximately 2,000 elements in an audio vector representation. Then the detection phase is applied by computing the average power signal in each 2,000 elements, followed by the normalization process. The variance is computed by considering 20 sub-frames.

Normally, our experiment should be conducted based on three (3) metrics which are - the false detection rate, the missed detection rate and the bad detection rate. The false detection rate reflects the number of times bird calls are identified in frame blocks, materialized by a positive flag whereas there were no bird call occurrence in the considered frame. The missed detection rate represents the proportion of missed detection flags considering the number of occurrence that are present in the frame block. Finally the bad detection rate refer to the percentage of bad detection flags, where the feature is detected tardily by the system. In our context, we focus on the first two metrics.

B. High SNR

Since the variance may be fluctuating while the signal-to-noise ratio gets lower, the determination of the variance threshold is dependent on the SNR. The most accurate values of the variance threshold that has been identified lie between 0.19 and 0.24. In the first experiment, we vary the signal-to-noise ratio from 0 to 35 dB and observe the behavior of the two metrics - missed and false detection rates.

Similar to that we were expecting, by referring to Figure 2, the missed and false detection rate are slightly constant. As a consequence an increase in the signal-to-noise ratio does not induce a variation of the proportion of missed and false detection flags. Even with 0 dB, a gain can already be observed in the sound quality, yet yielding acceptable likelihood for the bird call detection.

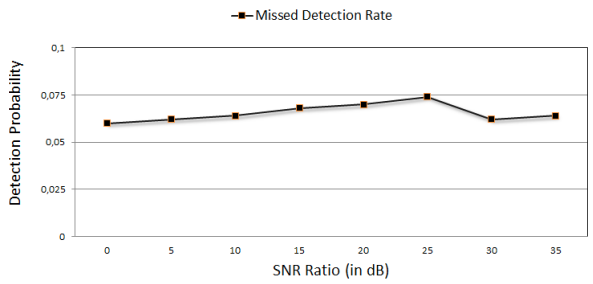
C. Reducing SNR

In the second experiment, we tested the ability of the protocol in providing satisfying results in the case of highly present noise. For this, we proceed by progressively decreasing the signal-to-noise ratio from 0 to -35 dB. The corresponding results are plotted in the Figure 2. In order to fit with the sensibility of the variance threshold considering the signal-to-noise ratio, we reduced the value of the variance threshold to 1.9. In this way, we take into account the unexpected fluctuation that might be caused by, for instances signal-to-noise ratio above -15 dB. After submitting to our scheme a series of data of approximately four hours, we plotted the evolution of the two metrics in Figure 2.(a) and 2.(b).

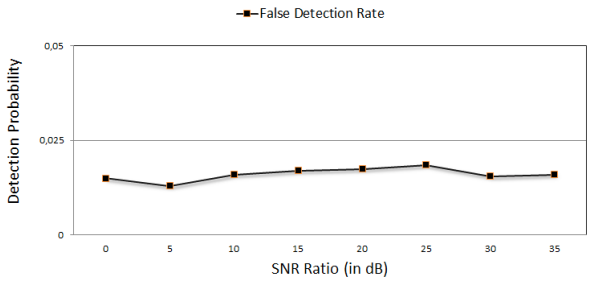
By observing the plotted results from the two boxes, it can be seen that the missed detection probability decreases sharply between the -35 and -20 dB, and falls into 30 % to reach 0 % between -15 and 0 dB. In the same time, no known evolution of the false detection rate is observed, by considering Figure 2.(b). The false detection rate fluctuates between 0.015 and 0.020. Then it can be conclude that the false detection rate is not influenced by the noise level. In another way, with high noise level inducing very low signal-to-noise levels, the probability of missing detection flags may increase.

V. Conclusion

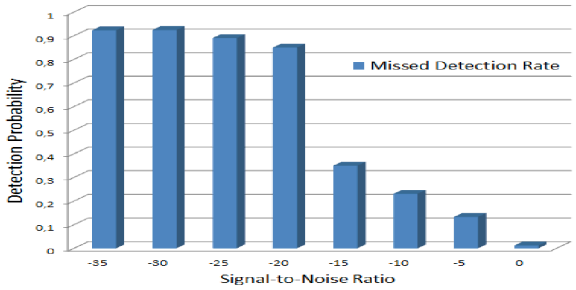
In this work, we aimed to study how bird calls can be detected by using normalized power sequences and manipulating computed variances, for the purpose of monitoring a rice field where grain-eating birds are likely to get introduced in. Considering



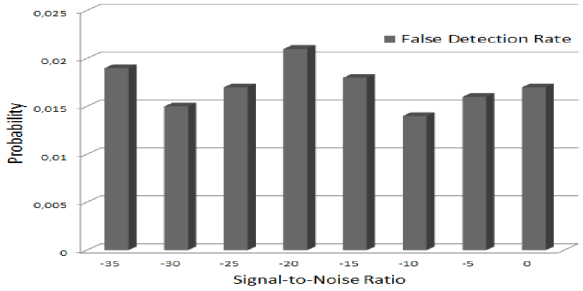
(a) Evolution of the missed detection rate with SNR above 0 dB.



(b) Evolution of the false detection rate with SNR from 0 to 35 dB.



(c) Evolution of the missed detection rate with SNR from -35 to 0 dB.



(d) Evolution of the false detection rate with SNR from -35 to 0 dB.

Fig. 2. Missed and False Detection Rate with SNR from 0 to 35 decibels.

it as the first research element in an automatic monitored system, obtained results show a detection percentage of about 91.07 % with relatively acceptable missed and false detection rate, by training the synthesized audio data set. Detecting harmonic sounds using statistical measures is not an evident task due to the complex structure of harmonic sounds. So the obtained results may reflect a little the data set composition. Our works in refining the proposed scheme are holding, and we aim to remodel the detection mechanism, eventually confront it with other audio feature extraction measures before applying it with real world deployed sensors.

REFERENCES

[1] Maximo Cobos, Juan J Perez-Solano, Santiago Felici-Castell, Jaume Segura, and Juan M Navarro. Cumulative-sum-based localization of sound events in low-cost wireless acoustic sensor networks. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(12):1792–1802, 2014.

[2] Selver Ezgi Kucukbay and Mustafa Sert. Audio-based event detection in office live environments using optimized mfcc-svm approach. In *Semantic Computing (ICSC), 2015 IEEE International Conference on*, pages 475–480. IEEE, 2015.

[3] Kwang Myung Jeon, Dong Yun Lee, Hong Kook Kim, and Myung J Lee. Acoustic surveillance of hazardous situations using nonnegative matrix factorization and hidden markov model. In *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

[4] Alain Dufaux. Detection and recognition of impulsive sounds signals. *Institute de Microtechnique Neuchatel, Switzerland*, 2001.

[5] CK Catchpole and PJB Slater. Bird song: biological themes and variations cambridge university press. *Cambridge, UK*, 1995.