WILDLIFE INTRUDER DETECTION USING SOUNDS CAPTURED BY ACOUSTIC SENSORS

*Marius Vasile Ghiurcau, Corneliu Rusu**

Technical University of Cluj-Napoca
Faculty of Electronics, Telecom. and Inf. Tech.
Baritiu 26-28, Cluj-Napoca, RO-400027, Romania

Radu Ciprian Bilcu

Nokia Research Center
Multimodal Sensing and Context
Visiokatu 1, 33720, Tampere, Finland

ABSTRACT

In this paper we classify the sounds originated from humans, birds and cars. The motivation of such a classification is to detect the intruders into protected wildlife regions such as protected forests, lakes, and other natural reservations. The proposed algorithm for sound encoding and classification is Time Encoded Signal Processing and Recognition (TESPAR) combined with the archetypes technique. We have tested our method on a database consisting of 300 recordings, 100 for each class, and several types of noise (white Gaussian noise, rain sound and wind sound) have been added to the recordings in order to simulate the different outdoor environments. Several pre-processing steps have been included and tested in order to verify the improvement of the classification accuracy. We performed a downsampling from 8 kHz to 6 kHz of the original recordings, followed by band pass filtering and the results shown an increased efficiency of TESPAR in the classification process.

Index Terms— sound classification, infinite clipping, TESPAR, LBG-VQ

1. INTRODUCTION

Humans play a critical role in ensuring the integrity of our forests and wild places. There are many natural reserves with wildlife, flora, fauna or features of geological or other special interest which are spread and there is practically impossible a continuous surveillance of all these areas. Although these regions are protected by law they are quite often the target of bad intentioned people for hunting, forest cutting and other. Moreover, the simple disturbance of wildlife by curious people could harm the endangered species. Not only the terrestrial reservations are the target of this illegal activities but also the protected lakes or coastal regions (such as the Danube Delta) are places of illegal fishing, or hunting of bird species strictly protected by international laws. At the same time, deforestation is proceeding at an unprecedented rate all over the world. The big problem is represented by the illegal deforestation which is hard to be detected in real time and stopped. Overall, systems of monitoring wild regions and detecting intruders are very necessary these days and would probably ensure a better preserving of these protected areas.

In this paper we shall evaluate the performances of a potential monitoring system that could be assimilated to an "acoustic eye". Its usage against other monitoring systems like the video surveillance ones, has some advantages: simplicity of implementation, less information to be processed, the fact that it does not depend on the visibility and a much cheaper solution. The ultimate goal of our work is to develop an acoustic sensor network which, placed inside a protected wildlife area, would be able to detect and classify several sounds of interest. The sounds of interest are related to several different events that must be monitored inside such protected areas. For the purpose of the work presented here we are interested in the detection and classification of only three sound classes: sounds originated from humans, birds and cars.

The problem of sound classification have been addressed many times in the open literature and several solutions have been proposed for different fields. Medical applications, like hearing aids and remote monitoring, have been addressed in [1]. Identification of the musical instruments from an audio recording represents an important field with applications to music retrieval for instance [2]. Environmental sound classification has also a wide range of applications and an important number of solutions have been proposed in the open literature [3]. The problem of vehicle identification have also been addressed, in [4], where the goal is to identify several types of vehicles based on the generated sound. Most of the standard classification (or recognition) of sounds methods imply extracting of different features.The question that may arise is when it must be realized a standalone system with very low complexity and low power consumption. Here we shall propose a solution that does not require many computations, and it could be easily implemented on a simple microcontroller.

This paper is organized as follows. At the beginning the theoretical backgrounds of this paper is presented in Section 2. Technical design and software implementation is provided in Section 3. Section 4 presents the results of our research and a short discussion. Some conclusions and future work possibilities can be found in Section 5.

2. THEORETICAL BACKGROUND

2.1. TESPAR Algorithm

TESPAR (Time Encoded Signal Processing and Recognition) is a simple and efficient language for describing complex waveforms in digital terms [5, 6], which is based on the infinite clipping theory of two researchers, Licklidder and Pollack. They investigated the effects of amplitude clipping on the intelligibility of the speech waveform. They managed to extend this process to the so-called infinite clipping format, where all the amplitude information was removed from the waveform. The result was a binary transformation that preserves only the zero-crossings points of the original signal. Using infinite clipping, mean random-word intelligibility scores of 97.9 % have been achieved.

Basically, the TESPAR method is based on counting the zeros of a digital signal. The simplest TESPAR coder uses two descriptors:

- D, which represents the duration between two consecutive real zeros (the number of samples between two real zeros);

*This research was supported by CNCSIS Grant ID 162/2008.

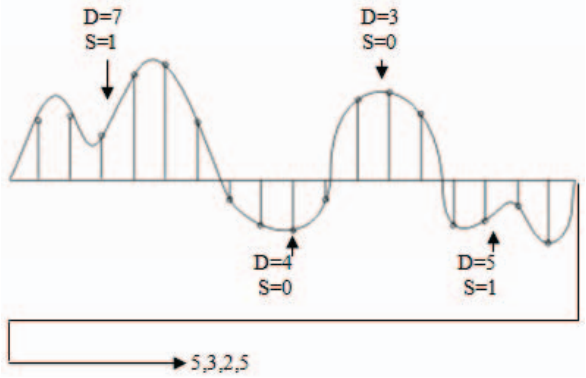

Fig. 1. TESPAR Symbol Stream.

- S, also called shape, is given by the number of points of local minima/maxima between two consecutive real zeros.

2.2. TESPAR Alphabet

The D/S pairs obtained from the analysis of the original audio signal are coded using an alphabet. This alphabet is the result of a vector quantization process [7, 8]. The result of the coding is a symbol stream similar to the one that can be seen in Figure 1.

2.3. TESPAR Matrixes

The symbol stream received from the TESPAR coder can be easily converted in a series of classification operands named TESPAR matrixes:

- $\mathbf{S}$ matrix: $N \times 1$ matrix-array vector that counts the number of apparitions for each symbol from the alphabet, in the coded symbol stream. Here N is the number of symbols from the alphabet.
- $\mathbf{A}$ matrix: $N \times N$ matrix, counts the number of occurrences of all the pairs of symbols, at a distance n (lag) apart; it contains more information than $\mathbf{S}$ matrix, but as a compromise, the computational time is larger.

2.4. Archetypes

Because of the fixed length of the TESPAR matrixes, one can use the archetypes technique in order to classify and recognize the class of a particular sound. The archetypes can be found by adding together and then computing the average of the $\mathbf{A}$ and $\mathbf{S}$ matrixes. The averaging process tends to emphasize the consistent characteristics of the condition and reduces the significance of anomalies that may exist in the individual examples [6]. Once the archetypes are computed, one can store them in databases and used them later on for classification of unknown samples. The process is simple, an $\mathbf{A}$ or $\mathbf{S}$ is computed and compared to each of the archetypes in the database. The closest is declared the winner [4].

3. DESIGN AND IMPLEMENTATION

3.1. Databases

For our research we have used three databases of over 100 recordings each.

- Database 1 is a small part of xeno-canto America [9], a database that contains over 25.000 recordings made by bird watchers from all over continental America. All the recordings are performed in different forests and areas with various species of birds.
- Database 2 was recorded by the authors and contains recordings of vehicle sounds (mostly sedan cars); this database was used also in other projects [4].
- Database 3 has over 100 recordings of speech sounds, most of them belonging to students from the Technical University of Cluj-Napoca. They were asked to record themselves when uttering their names or different sentences.

Especially for databases 2 and 3, the recordings are quite short, 2-4 seconds each. This is quite close to real situations, when for example a car passes through a "check point", where an acoustic sensor is placed and the useful recorded sound will not exceed 2-3 seconds.

3.2. Preprocessing stage

The recordings from the databases had to be preprocessed before going further. First step was to downsample all of them at a sample frequency of 8 kHz. For this Sony Sound Forge 7.0 software was used. As a result, we obtained a database of over 300 16-bits mono audio *.wav files. We have to note a particularity of the TESPAR algorithm: it very sensible to the DC offset, so we decided to remove this component from the signals. Using MATLAB, we computed the mean values of each signal and subtract it from every available sample.

3.3. MATLAB implementation

For the current research, all the software was implemented under MATLAB platform. The steps of the implementation process were the following:

3.3.1. Step 1

Firstly, the script that is analyzing the *.wav files and computes the values for D and respectively S between every two consecutive real zeros was implemented. A simple computation led us to the following two conclusions: 97.8% of the values for D were situated in the [0, 40] interval and around 97.7% of the values for S were situated in the [0, 8] interval. Consequently, in order to decrease the size of the alphabet used later on in the encoding process and also for a better approximation of the signals, we decided to limit the values for D and S at 40 and respectively 8. Everything exceeding 40 respectively 8 was set to these two values.

3.3.2. Step 2

We have implemented the Linde-Buzo-Gray [7, 8] vector quantization algorithm in order to generate the alphabet. For this computation a total of about 80 seconds of recordings from birds, humans and cars were used (around 100.000 D/S pairs). A 32 symbols alphabet was considered to be more than enough for our research.

3.3.3. Step 3

The modules that generate the TESPAR symbol stream and the TESPAR matrixes were implemented. Figures 2 and 3 present examples of TESPAR $\mathbf{S}$ and respectively $\mathbf{A}$ matrixes. We have also implemented the scripts that compute the archetypes for both $\mathbf{A}$ and $\mathbf{S}$

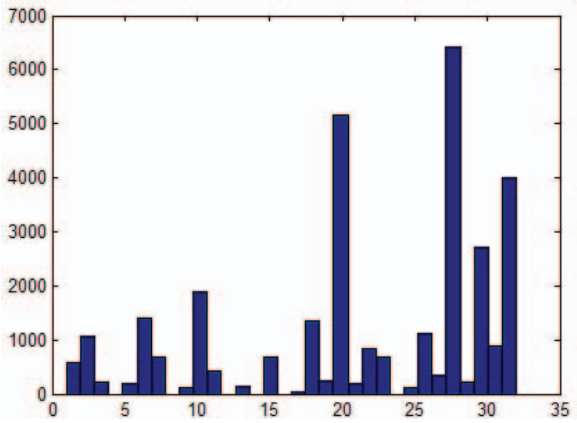

Fig. 2. Example of **S** matrix.

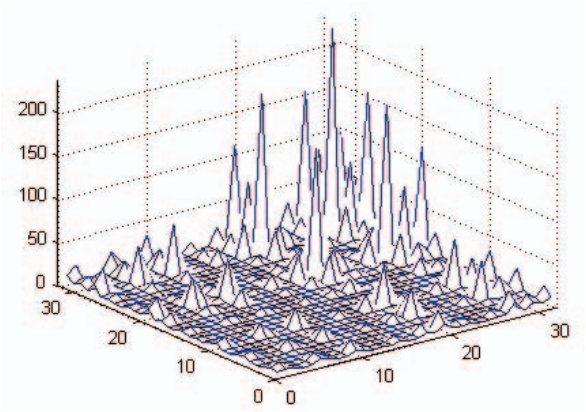

Fig. 3. Example of **A** matrix.

matrixes. For the **A** matrix the value for the lag was set to $n = 2$. In all the cases (birds, humans, cars) we have used 10 utterances, different from the test utterances, in order to generate the archetypes. The archetypes were finally stored in a database. In order to get better statistically results, for each experiment we have decided to generate ten archetypes, using ten different utterances each time and use for testing the remaining sounds that were not used for training.

3.3.4. Step 4

Last step was to assemble together all the modules that were implemented and test the final software for all the test sounds that were available. As it was said before, each time an unknown sound sample appears, its corresponding **S** and/or **A** matrixes are computed and then compared with the archetypes stored in the databases. First a normalization of these matrixes is needed because the different lengths of the sound samples imply different lengths for the symbol streams that generate the archetypes and also the new matrixes. The comparison is done using city block distance, the corresponding archetype of the smallest distance being declared the "winner".

3.3.5. Step 5

After the initial testing we decided to perform a band pass filtering of the signals and then to check again the classification rates, to see if any improvement is possible. For this we have used 10th order band pass Butterworth filters with the low cut off frequency of 70Hz and high cut off frequency of 3000 Hz. We have tried this band-pass filtering in order to remove some of the noise present in the recordings and also for discarding possible unwanted information. The noise in the recordings comes from each particular recording device, which obviously, is introducing noise when performing a recoding. Also, the high-pass filtering with the 70 Hz was used in order to prevent the eventual noise from the interference of the recording devices with the electrical wiring network which works at 50/60 Hz.

S matrix	human	bird	car
human	97.9	0	2.1
bird	0	93.3	6.7
car	5.2	0	94.8

Table 1. Confusion matrix for **S** matrix

A matrix	human	bird	car
human	98.2	0	1.8
bird	0	93.8	6.2
car	0	0	100

Table 2. Confusion matrix for **A** matrix

3.3.6. Step 6

Finally, we decided to perform a few more experiments, in order to get results as close as possible to real environment conditions:

- add additive white Gaussian noise (AWGN) over the test recordings, with a SNR of 20 dB;
- add rain and wind sounds (real sounds, recorded in the nature [10]) over the test recordings; maximum amplitude of the noise sounds was set to be $1/3$ of the maximum amplitude of the clear sounds;
- downsample the initial recordings.

One of the reason for the downsampling was to verify how much would affect this our classification rates; our final goal is to use this application mostly in wildlife places, on low cost systems for which any reducing of the amount of data to be processed would be significant (downsampling from 8 kHz to 6 kHz would reduce approximately with 25% the amount of data). This downsampling requires us to redo all the previous steps, and so a limitation of the values for D at 35 and for S at 8 was set. We recomputed the archetypes for S and A matrixes and performed again the tests. The results are presented in Section 4.

4. RESULTS AND DISCUSSIONS

4.1. Experimental Results

Table 1 presents the results of the classification experiment when matrix **S** is used; Table 2 presents the results of the classification experiment when matrix **A** is used (*Experiment 1*).

The above results can be slightly improved, by the band-pass filtering process, as it can be seen in Tables 3 and 4 (*Experiment 2*).

Besides the above mentioned base-line tests we have done several other experiments in order to verify the performance of the system in several more restrictive situations. For this purpose the following experiments have been done:

S matrix	human	bird	car
human	100	0	0
bird	0	95.5	4.5
car	2.2	9.3	88.5

Table 3. Confusion matrix for **S** matrix with band-pass filtering performed

A matrix	human	bird	car
human	100	0	0
bird	0	98.3	1.7
car	0	2.3	97.7

Table 4. Confusion matrix for **A** matrix with band-pass filtering performed

Experiment 3: downsampling of the initial recordings from 8 kHz to 6 kHz led us to 95.3% correct classification rate for both **S** and **A** matrixes. We also tried downsamplig at 4 kHz and the classification rates were situated around 88%-90%.

Experiment 4: we reduced the size of the coding alphabet from 32 symbols to 16 and the classification rates received were 90% for S matrix and around 98% for the **A** matrix.

Experiment 5: adaptive white Gaussian-distributed noise have been added to the test recordings and also band-pass filtering was employed. The correct classification rate for the **S** matrix was 74% and 84% for the **A** matrix. We also increased the number of symbols from the alphabet, from 32 to 64, but this improved our overall results with less than 1%.

Experiment 6: rain sound added to initial recordings (plus band-pass filtering). Our results show a 74.6% rate for **S** matrix and 89.3% for **A** matrix.

Experiment 7: wind sound added to initial recordings. For the **S** matrix we had a 71.3% rate and for the **A** matrix 88.6%.

4.2. Discussion

We can notice that in the case of using the **S** matrix, the overall correct classification rate is 95.3% and in the case of using A matrix is 97.3%. As it was expected, when using **A** matrix the correct classification rate is higher. This is because matrix **A** is supposed to contain more information than **S** matrix; it is obvious that for this particular situation the use of the **A** matrix would be more suitable. Another important aspect is that the humans are apparently easier to be recognized than cars or birds; for both **A** and **S** matrixes in the case of human sounds the classification rates are the highest of all.

Band pass filtering increases slightly the results (1.3%), in the case of **A** matrix (94.6% for **S** matrix and 98.6% for **A**). Downsampling the recordings and consequently lowering the amount of data to be processed seems to decrease the classification rates with approximately 2 to 3%.

When using clear sounds it seems that reducing the size of the alphabet to only 16 symbols decreases the rates with around 1% for **A** matrix and 5% for **S** matrix. In order to reduce the amount of computations, for the future it would be interesting to try using an alphabet with only 20-24 symbols.

One can notice the classification rates are situated between 85%-90% in the situation when sounds, affected by different types of noises (rain, wind, AWGN), are used for testing. For our purpose these results are quite good even though we have to admit that some improvements are needed.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have evaluated the performances of a potential monitoring system that could be assimilated to an "acoustic eye". An important conclusion that has to be pointed out is that TESPAR seems to be efficient in this classification of sounds originated from humans, birds and cars. Our correct classification rates of over 94% when using the **S** matrix and over 98% when using **A** matrix are notable. When the recordings are altered by noise, the rates decrease and an improving of these rates would be good. The usage of the same recoding device for the test and training sounds might improve our rates (only database 2 contains recordings with the same microphone, all the others are made with unknown microphones or any other recording devices by different persons).

Another important aspect is the usual length of the recordings, which exceeds 2 to 4 seconds just in few cases. It is known that this influences significantly the classification rates when using standard methods. A future goal will be a comparison of the classification rates when using the combination of standard methods and TESPAR. Moreover, we intend to extend our databases with other particular sounds, like motorboats, chainsaws or heavy truck cars. Finally a hardware implementation of this project is desired.

6. REFERENCES

[1] Lucas Cuadra, Enrique Alexandre, Lorena Alvarez, and Manuel Rosa-Zurera, "Reducing the computational cost for sound classification in hearing aids by selecting features via genetic algorithms with restricted search," in *Proc. IEEE ICALIP 2008*, Shanghai, China, Jul 2008, pp. 1320–1327.

[2] Emmanouil Benetos, Margarita Kotti, and Constantine Kotropoulos, "Musical instrument classification using non-negative matrix factorization algorithms," in *Proc. IEEE ISCAS 2006*, Kos, Greece, 2006.

[3] Byeong-jun Han and Eenjun Hwang, "Environmental sound classification based on feature collaboration," in *Proc. IEEE ICME 2009*, New York, USA, Jun 2009, pp. 542–545.

[4] Marius Vasile Ghiurcau and Corneliu Rusu, "Vehicle sound classification application and low pass filtering influence," in *Proc. IEEE ISSCS 2009*, Iasi, Romania, July 9–11, 2009, pp. 301–304.

[5] Hydralogica, "Digital signal processing solutions," http://www.hydralogica.com/technology.htm.

[6] R. A. King and T. C. Philips, "Shannon, TESPAR and approximation strategies," *Computers and Security*, vol. 18, pp. 445–453, 1999.

[7] Cornel Balint, "An improved LBG algorithm for vector quantization," *Annals Computer Science*, vol. 2, http://www.anale-informatica.tibiscus.ro/download/lucrari/2-1-02-Balint.pdf.

[8] J. C. Licklidder and I. Pollack, "Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech," *J. Acoust. Soc. Am.*, vol. 20, pp. 42–51, Jan. 1948.

[9] Xeno-canto, "Bird sounds from America," http://www.xeno-canto.org/.

[10] Audio4fun.com, "Free download nature sounds," http://funny-stuff.audio4fun.com/sound-effects.php?page=nature-sounds&id=21436.

[11] Data-Compression, "Vector quantization," http://www.data-compression.com/vq.html.