Towards an application for detecting intruders in wildlife regions

TOWARDS AN APPLICATION FOR DETECTING INTRUDERS IN WILDLIFE REGIONS

Marius Vasile Ghiurcau[*], *Corneliu Rusu*[*], *Jaakko Astola*[**], *Radu Ciprian Bilcu*[***]

[*]Signal Processing Group, Faculty of Electronics, Telecommunications and Information Technology
Technical University of Cluj-Napoca
Cluj-Napoca, Str. Baritiu 26-28, RO-400027, Romania
[**]Tampere International Center for Signal Processing, Department of Signal Processing
Tampere University of Technology
P.O. Box 553 FI-33101, Tampere, Finland
[***]Nokia Research Center
Multimodal Sensing and Context
Visiokatu 1, 33720, Tampere, Finland

ABSTRACT

Sound classification is a topic that has been of a major interest for the scientific community. Recently, a low complexity solution for classifying sounds in wildlife regions has been proposed. The motivation of such a classification was in detecting the intruders from these regions. In this paper we propose a different approach, one that uses Mel-frequency cepstral coefficients in a Gaussian Mixture Model framework. The tests are performed on 4 databases of 100 recordings each. The sounds of interest are represented by recordings from humans, cars, birds and animals. In order to simulate situations as close as possible to real environments, several types of noises have been considered. The new approach proves to be more robust than the previous one at the cost of increased computational complexity. Since low complexity systems are more likely to be feasible for wildlife applications, the complexity issue is discussed and a solution is proposed.

1. INTRODUCTION

Sound classification shows a continuous development and its applicability was demonstrated in various fields. Besides the well-known speech/speaker recognition applications, there are also other areas in which sound classification proved to be successful. Medical applications, like classification of heart sounds [1], hearing aids [2] or remote monitoring systems [3] are very popular these days. Different solutions for environmental sound classification applications were proposed in [4, 5]. An application for the classification of acoustic events in a kitchen environment can be found in [6]. Also, vehicle identification using wireless sensor networks [7] is a promising topic, with different applications in real life.

Natural reserves, forests, protected lakes, or the coastal regions, are very often the target of different intruders interested in forest cutting, hunting, or simple curious people. The actions of these intruders could affect the continuity of the endangered species and the overall integrity of the wild places. As a consequence, systems for monitoring these wild regions are needed. Video surveillance systems can not be considered as a unique solution for such purposes. The main reason is related to their increased complexity, amount of information that has to be processed, high power consumption and of course their high costs. We want to propose a different approach for this type of surveillance system, one that uses sounds not image. The final goal is to realize an acoustic sensor network that could be seen as an 'acoustic eye'.

In recent studies [8, 9] a low complexity solution for detecting intruders in wildlife regions has been proposed. The sound classification algorithm used was based on Time Encoded Speech Processing and Recognition (TESPAR). The most simple TESPAR coder is using two descriptors for each segment situated between two consecutive zeros of a signal:

- D - number of samples between two consecutive zeros;
- S - number of points of minima/maxima in the segment.

The D/S pairs are used to encode the original signal using an alphabet. This alphabet is the result of a vector quantization process. Basically the resulted symbol stream is converted in some classification operands called TESPAR matrices (**A** and **S** matrices). Finally, some archetypes are constructed for each class of sounds, and they are used later in the classification process. More details about TESPAR can be found in [10].

The work in [9] has improved the results from [8], especially when the **S** matrices have been used in the classification process. This proved to be very important; indeed, using only the **S** matrices, this would lead to a decrease in the complexity of the algorithm, which can be crucial in a standalone system with low power consumption. Even though there was also an improvement in the classification rates when various types of noisy environments were simulated, the rates were not fully satisfying.

In this paper we propose a sound classification approach that uses Mel-frequency cepstral coefficients in a Gaussian Mixture Model framework. The motivation of this work is threefold:

1. We want to compare the results of these two different approaches, to see exactly how well does TESPAR perform in comparison to standard sound classification methods.
2. Taking into consideration the improved results of the standard classifier presented in this work, we suggest a combined solution, which utilizes both of these approaches.
3. Finally, even though our purpose is to realize a system that has to be used in wildlife, our method can also be used for property surveillance. In this different situation low power consumption should not be mandatory anymore, thus more complex and robust algorithms can be utilized.

The rest of the paper is organized as follows. The theoretical background of this paper is presented in Section 2. Technical design and implementation details are provided in Section 3. Section 4 presents the experimental results and Section 5 concludes the paper.

2. THEORETICAL BACKGROUND

2.1 Mel-Frequency Cepstral Coeficients Overview

One of the most popular features used in sound classification applications are the Mel-frequency cepstral coefficients (MFCCs). They are a short-term spectrum-based feature which give good discriminative performance.

The extraction of the MFCCs includes the following steps:

1. Pre-emphasis: for reducing the noise and also for enhancing high-frequency spectrum, a finite-order impulse response (FIR) filter is applied to the audio signal:

$$H_{\text{FIR}}(z) = 1 - az^{-1}$$

The value for a is usually selected from the [0.95, 0.98] interval.

2. After the pre-emphasis, the signal is divided into frames. This framing comes from the necessity of transforming the signal into statistically stationary blocks. Overlapping frames with a 30-50% overlap are used, in order to avoid losing information at the end of the frames.
3. For preventing abrupt changes at the end points of the frames, a window function is used (usually a Hamming window):
4. For each frame the Discrete Fourier Transform (DFT) is applied. Because humans do not perceive pitch linearly, the frequency band has to be divided using a filter-bank of triangular filters spaced on the Mel-scale [11]:

$$\text{Mel}(f) = 1127 \ln\left(1 + \frac{f}{700}\right)$$

5. Spectral envelope in dB is obtained by applying logarithm to the amplitude spectrum. Finally, the discrete cosine transform (DCT) is applied [12]:

$$c_n = \sum_{j=1}^{M} \log S_j \, cos\left[\frac{\pi n}{M}(j - \frac{1}{2})\right], \quad n = 1, 2, ..., N,$$

where
- c_n is the n^{th} MFCC coefficient;
- M is the number of filterbanks;
- N is the number of coefficients one wants to compute;
- S_j is the magnitude response of the j^{th} filterbank channel.

The zeroth coefficient is usually dropped because it is the average log-energy of the frames. Most of the times first and second order differences of the MFCCs are included as a feature. Those are called delta and delta-delta coefficients.

2.2 Gaussian Mixture Models

A Gaussian Mixture Model can be written as a weighted sum of M component densities and has the following form [13]:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x)$$

where x is a d-dimensional random vector, $p_i(x), i = 1, ..., M$, is the component density and $w_i, i = 1, ..., M$, is the mixture weight.

The component densities are d-variate Gaussian functions given by [13]:

$$p_i(x) = \frac{1}{\sqrt{(2\pi)^d det(\Sigma_i)}} \exp\left[-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right]$$

where μ_i is the mean, Σ_i is the covariance matrix, d is the number of features incorporated into every feature vector.

The weights w_i have to satisfy the following relation: $\sum_{i=1}^{M} w_i = 1$. Each model can be written as a function of the following parameters: $\lambda = (w_i, \mu_i, \Sigma_i)$, $i = 1, ..., M$.

The log of the likelihood function is [14]:

$$\ln p(x|w, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{i=1}^{M} w_i p\left(x_n|\mu_i, \Sigma_i\right) \right\}$$

where $x = x_1, ..., x_N$

Finally, for finding the maximum likelihood solutions for the models different algorithms are used. In the present study, the expectation-maximization (EM) algorithm was employed, which is known as an elegant and powerful method for finding a maximum likelihood solution [14, 15].

3. DESIGN AND IMPLEMENTATION

3.1 Databases

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

- Database 1 is a small part of xeno-canto America, 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.
- 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.
- Database 4 has contains recordings of different animal sounds: lions, bears, snakes, horses, cows, cats, frogs and others. All the recordings were collected from different animal databases on the internet.

In the previous studies [8, 9] only the first three databases were used. In this approach we present some comparative results with the first study and also the new results after Database 4 is introduced.

3.2 Experimental setup

Our practical work is structured as follows. Firstly, the Mel-frequency cepstral coefficients were extracted from the signals and then the training of the models and testing using the

	Number of MFCCs		
GMMs	10	14	20
10	97.33	98.00	99.33
30	98.66	99.00	99.66
50	99.00	99.00	99.66

Table 1: Classification rates for clean sounds (human, car, bird)

	Number of MFCCs		
GMMs	10	14	20
10	84.66	85.33	92.00
20	85.00	91.00	94.00
30	88.66	91.66	95.00
40	88.00	92.00	96.66
50	90.00	94.66	97.33

Table 2: Classification rates when rain is added to the test sounds (human, car, bird)

Gaussian-Mixture Models was employed. For each class of sounds we had 100 recordings of a few seconds each. All the recordings were sampled at 8 kHz and stored in 16 bits mono *.wav files.

For pre-emphasis a (FIR) filter with the pre-emphasing coefficient *a*=0.97 was used. The signal was divided in 256-sample frames with an overlap of 128 samples (50% overlap). This corresponds to a frame length of 32 ms. We also tried decreasing of the frame length, but the results shown no improvement. A filter-bank of 40 triangular filters spaced on the Mel-scale was used.

In our experiments we have implemented 10 to 20 cepstral coefficients and the results were compared. First coefficient was all the times discarded, as it is dependent of the channel gain. Along with MFCCs we have also used the delta coefficients.

For modeling the Gaussian Mixtures, different numbers of Gaussian components were selected. The expectation maximization algorithm was used, with a maximum of 1000 iterations and the value for the threshold was set to 0.01.

In order to get statistically good results, when performing our tests we have considered *leave one out cross-validation* [16] method. Consequently, we separated the sounds in two sets: testing set and validation set; 99 recordings were used for training the system and the remaining recording was used for testing. This process has been repeated 100 times, every time changing the test recording.

After the regular experiments were performed, we decided to simulate some scenarios that could be encountered in real life. For this, noise was added to the test signals. The training of the models was done with clean sounds, while for testing we added rain and wind sounds (real sounds, recorded in the nature) to the test signals. The maximum amplitude of the noise signals was set to 1/3 of the maximum amplitude of the clean recordings in the databases.

4. RESULTS AND DISCUSSIONS

Experiment 1: In the first experiments we tried to compare the results from our previous studies with the results obtained in this new approach. Table 1 presents the results when the sounds of interests where only the birds, humans and cars, and clean sounds where used for both training and testing. We also tried adding noise to the test signals, in order to see how this influences the classification rates. The results with rain added to the test sounds are presented in Table 2. When wind was added to the test signals the best correct classification rate obtained was 93.33%.

	Number of MFCCs			
GMMs	10	14	18	20
10	92.75	93.75	93.25	92.75
20	95.25	94.50	95.00	96.00
30	94.25	96.00	95.25	96.25
40	94.75	95.50	96.00	96.25
50	95.25	96.50	96.00	95.50

Table 3: Classification rates for clean sounds (human, car, bird, animal)

	Number of MFCCs			
GMMs	10	14	18	20
10	74.75	79.25	79.75	78.25
20	80.25	81.75	82.50	84.00
30	81.50	84.00	82.50	82.25
40	81.25	83.25	83.00	85.00
50	81.25	82.75	84.75	86.50

Table 4: Classification rates when rain sound is added to the test sounds (human, car, bird, animal)

Experiment 2: We repeated the previous tests but with an extra class of sounds of interest, represented by animals. The results for this study are presented in Table 3 and the confusion matrix for MFCC=20 and GMM=40 can be seen in Table 5.

Experiment 3: We added again rain to the test sounds (all four databases), similar to the procedure explained in *Experiment 1*. Table 4 shows the evolution of the classification rates in this case. For the combination MFCC=20 and GMM=50 the confusion matrix can be observed in Table 6. When wind was added to the test sounds, our classification rates were situated between 70 and approximately 80%; best score was 82.25%.

The results from *Experiment 1* show us an increase from our previous approach [9]. For clean sounds we managed to achieve an overall correct classification rate of 99.66% while previously our best score was 97.33%. When the test sounds were affected by noise our scores were 97.33% for rain and 93.33% for wind. Previously we had 94% for rain and 89.33% for wind. Obviously, the standard classifier performs better than TESPAR.

When using the forth database also, we encountered a decrease in the classification rates. For clean sounds, the best scores obtained were of 97.33%. As it can be seen in the confusion matrix from Table 6 for both human and car we have 100% correct classification rate. After adding noise to the test sounds, the decrease was considerable. Best score obtained was 86.50% for rain and 82.25% for wind sounds.

Even though, in the confusion matrix presented in Table 6 an interesting aspect can be noticed. For human we have a correct classification rate of 98%, while for car it is 97%, which can be considered quite satisfying. The worst results are for animals, where only 57% were correctly classified. This results was somehow expected, since our databases with

	bird	car	animal	human
bird	94	0	6	0
car	0	100	0	0
animal	9	0	91	0
human	0	0	0	100

Table 5: Confusion matrix, clean sounds

	bird	car	animal	human
bird	94	2	4	0
car	3	97	0	0
animal	43	0	57	0
human	0	0	2	98

Table 6: Confusion matrix, rain sound added

animals contains various species, so the classifier can not really construct an accurate model, because the sounds present are not very similar.

Another interesting aspect that has to be pointed out here is that all the misclassified animals were considered as birds. For the purpose of our goals, this aspect does not affect us too much. Indeed, we are interested in intruders, namely cars or human; if an animal is considered a bird or vice versa, that may be rather acceptable for a first step in analysis.

5. CONCLUSIONS AND FUTURE WORK

As one may expect, the standard sound classification method presented in this paper proved to be more robust than the low complexity solution suggested in the previous works. However, we are aware that such a complex system could not be implemented easy on a cheap controller and placed in a wildlife region. Even though, a possible combined solution that overcomes this difficulty will be tried.

A future goal is to develop a low complexity system that identifies possible intruders and sends to a base station the corresponding recording. At the base station, one can try more complex approaches in order to make sure that we are facing with an intruder.

Moreover, an intruder verification system seems to be more suitable for our goals. Consequently, because of the various sounds encountered in the nature, we would think of a slightly different approach, in which the low complexity system does not try to classify the sounds in different classes, but only only checks if a certain recorded event belongs to a human, a car or an engine, a gun shot or other possible sounds of interest that could be considered as an intruder.

A higher threshold could be set, even though this could lead to the possibility of increasing false alarms. Obviously, a certain compromise has to be made, when setting the threshold, because a high number of false alarms could lead to a 'system jam'.

Finally, one of the future goals is to increase our databases, with sounds that reproduce gun shot, thunder, chain saws etc.

ACKNOWLEDGEMENT

The work of Marius Ghiurcau was supported by PRODOC POSDRU/6/1.5/S/5 ID 7676. The work of Corneliu Rusu was supported by CNCSIS Grant ID 162/2008.

REFERENCES

[1] H. M. Hadi, M. Y. Mashor, M. S. Mohamed, and K. B. Tat, "Classification of heart sounds using wavelets and neural networks," in *Proceedings of IEEE CCE 2008*, Mexico City, Mexico, Nov 2008, pp. 177–180.

[2] E. Alexandre, L. Cuadra, M. Rosa, and F. Lopez-Ferreras, "Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms," in *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15. no. 8, pp. 277-288, Nov. 2007.

[3] D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J.-F. Serignat, "Information extraction from sound for medical telemonitoring," in *IEEE Trans. on Information Technology in Biomedicine*, vol. 10. no. 2, pp. 264-274, Apr. 2006.

[4] B. Han and E. Hwang, "Environmental sound classification based on feature collaboration," in *Proceedings of IEEE ICME 2009*, New York, USA, Jun 2009, pp. 542–545.

[5] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time-frequency audio features," in *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17. no. 6, pp. 1142-1158, Aug. 2009.

[6] F. Kraft, R. Malkin, T. Schaaf, and A. Waibel, "Temporal ICA for classification of acoustic events in a kitchen environment," in *Proceedings of IEEE Interspeech 2005*, Lisbon, Portugal, Sep 2005, pp. 2689–2692.

[7] S. S. Yang, Y. G. Kim, and H. Choi, "Vehicle identification using wireless sensor networks," in *Proceedings of IEEE SoutheastCon 2007*, Richmond, USA, Mar 2007, pp. 41–46.

[8] M. V. Ghiurcau, C. Rusu, and R. Bilcu, "Wildlife intruder detection using sounds captured by acoustic sensors," *Proceedings of IEEE ICASSP 2010*, Dallas, USA, 2010.

[9] ——, "A modified TESPAR algorithm for wildlife sound classification," *Proceedings of IEEE ISCAS 2010*, Paris, France, 2010.

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

[11] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-IEEE Press, 1999.

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK V3.1)*. Cambridge University, 2000.

[13] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," in *IEEE Trans. Speech and Audio Processing*, vol. 3, no. I, pp. 72-82, 1995.

[14] C. Bishop, *Pattern Recognition and Machine learning*. Springer Science, New York, 2006.

[15] G. McLachlan and T. Krishnan, *The EM Algorithm and its Extensions*. John Wiley&Sons, New York, 1997.

[16] R. Duda, P. Hart, and D. Stork, *Pattern Classification (2nd edition)*. Wiley-Interscience, 2000.