DETECTION OF HUMAN ACTIVITIES IN NATURAL ENVIRONMENTS BASED ON THEIR ACOUSTIC EMISSIONS

Stavros Ntalampiras$^{\pm}$*, Ilyas Potamitis*†

$^{\pm}$ Dep. of Electronics & Information, Politechnico di Milano, Milano, Italy
† Technological Educational Institute of Crete, Dep. of Music Technology and Acoustics, Crete, Greece

ABSTRACT

In this work we address the problem of detecting human activities in natural environments based solely on the respective acoustic emissions. The primary goal is the continuous acoustic surveillance of a particular natural scene for illegal human activities (trespassing, hunting etc.) in order to promptly alert an authorized officer for taking the appropriate measures. We constructed a novel system which is mainly characterized by its hierarchical structure as well as the variety of the acoustic parameters. Each sound class is represented by a hidden Markov model created using descriptors from the time, frequency and wavelet domains. We conducted extensive experiments for assessing the performance of the system with respect to its recognition and detection capabilities.

Index Terms— Generalized sound recognition, sound event detection, nature preservation, situation management and analysis.

1. INTRODUCTION

Natural environments are of critical importance with respect to the modern society for social as well as economic reasons. The living world provides extremely useful types of resources, many of which are necessary for preserving today's quality of life. The most useful types are: resource for food supply, energy source, particularly major source of medicines and huge source of raw materials which are primarily used by the industry. In this context, humans are given a plethora of choices as regards to their current needs by the diversity of nature. Moreover, this spectrum of resources enhances the role of nature towards offering solutions with respect to future needs and challenges of mankind.

Human activities are crucial as regards to preserving the integrity of our forests and wild places. Several natural reserves present properties which are very interesting as regards to their fauna, flora, geology etc. Continuous monitoring of the specific regions would be of great usage towards their protection as well as indexing biodiversity. However the cost of all day supervised scene monitoring is prohibitive and autonomous unattended supervision can be of great assistance towards creating a safe environment which benefits the biological diversity. Furthermore some areas of interest may be difficult or even impossible to access. Even though the majority of these places are protected by law (such as the Hymettus mountain located in the area of Attica, Greece), several human activities which may be harmful to the environment often take place (e.g. hunting, forest-cutting etc.). Additionally, some species may be harmed by even slight disturbances coming from inquisitive people. Similar situations apply to lakes or coastal regions where illegal fishing or bird hunting takes place. Illegal forest-cutting is one of the main causes of the rapid deforestation that has been observed during the recent years and it presents multiple societal and environmental problems while its immediate detection is of particular importance in order to limit the consequences as much as possible. The long-term implications of global deforestation are almost certain to jeopardize life on Earth as we know it. Other implications include loss of biodiversity, elimination of forest-based societies and climatic disruption. For all these reasons automatic monitoring of the natural environments has become a necessity since it can be proven helpful towards preserving them.

We suggest that systematic and non-intrusive, acoustic control can be carried out with remote monitoring stations in order to detect human presence. In this work *audio surveillance* includes the continuous capturing the audio sequence of a particular space and processing it for detecting sound events which are indicative of human activities. This definition clearly states that audio surveillance constitutes a branch of the generalized sound recognition technology. The closest paper to this work is [1], where sounds originated from humans (speech), birds and cars are considered in order to detect intruders. They employed the Time Encoded Signal Processing and Recognition (TESPAR) algorithm while they used a database of 300 recordings. The classification is based on the archetypes technique, where a comparison is made between the novel and the already "seen" patterns. The class of the archetype with the smallest distance is assigned to the

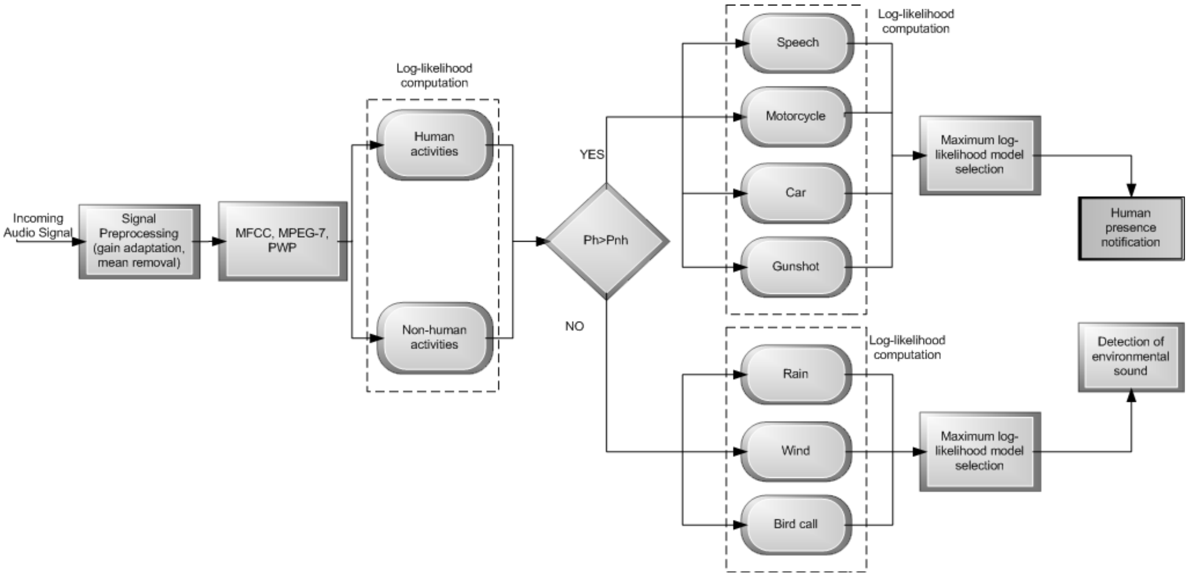
Incoming Audio Signal
Signal Preprocessing (gain adaptation, mean removal)
MFCC, MPEG-7, PWP
Log-likelihood computation
Human activities
Non-human activities
YES
NO
Ph>Pnh
Log-likelihood computation
Speech
Motorcycle
Car
Gunshot
Maximum log-likelihood model selection
Human presence notification
Log-likelihood computation
Rain
Wind
Bird call
Maximum log-likelihood model selection
Detection of environmental sound

Figure 1: *The block diagram of the two-stage topology for detection of human presence in natural environments.*

novel audio signal. Moreover, their experiments included addition of three types of noise (Gaussian noise, rain sound and wind sound) while the results are in the form of confusion matrices.

Our work is based on the experience gained from the scientific field of acoustic surveillance of hazardous situations. An exhaustive survey of the approaches that have been followed with respect to the particular field can be found in [2]. The emphasis of previous approaches is mainly placed on the classifier, the feature extraction process, the training data and the number of classes. Due to the similarity between these two problems, which essentially include *detection of sound events*, we adapted the specific technology to the one of acoustic detection of human presence in natural environments. The block diagram of the proposed sound recognition system is depicted in Fig, 1. We aim at detecting sound events related to manifestations of human activities (*speech*, *gunshot*, *car* and *motorcycle* sound events) in natural reserves based on the signal captured by a single microphone. Three types of non-human sound events are considered: *bird call*, *rain* and *wind*. Our target is to provide a generic methodology that a) leads to high detection accuracy, b) is easily adaptable to related problems and c) is practical (fast to set-up and operate). The ultimate goal of the proposed framework is to alert the authorized personnel regarding illegal human presence in order to take appropriate measures so as to minimize any type of potential catastrophic consequences. We employ a multi-domain characterization of the involved audio signals with parameters that belong to time, frequency and wavelet domain. Furthermore our dataset is thorough and concise after combining several well documented professional sound effect collections which contain audio of high quality.

The rest of this paper is organized as follows: in Section 2 we provide a short analysis with respect to each module of the system. Section 3 describes the stages of the development of the final system while Section 4 explains the experimental procedure where simulations of human activities in natural environments were artificially created. Our conclusions are drawn in the last section.

2. THE MODULES OF THE SURVEILLANCE FRAMEWORK

Acoustic features

In this paragraph we mention the groups of descriptors that were employed for the aims of this work. We have basically employed the following three groups of acoustic parameters:

a) Mel Frequency Cepstral Coefficients (MFCC)
b) MPEG-7 Audio Standard Low Level Descriptors [3]
c) Perceptual Wavelet Packet integration (PWP) [4]

These three groups are firstly used separately in order to assess their effectiveness with respect to the specific task. A series of experiments was designed so as to select the combination of parameters which provides the highest recognition rates. Parameters which belong to different domains may provide improved recognition accuracy when compared to any kind of mono-domain representation since they tend to capture diverse aspects of the structure of the audio signal. It should be mentioned that during the last phase of the computation of the MPEG-7 Audio Standard LLDs as well as PWP features, the DCT is used in order to

Table 1. *Two stage topology recognition rates (%) using the multi-domain feature set with respect to the sound classes related to human activities (average rate=79%, HMMs with 3 states and 16 Gaussian mixtures).*

Responded / Presented	Motorcycle	Car	Speech	Gunshot
Motorcycle	**70,1**	24,3	0	5,6
Car	20,4	**75,8**	0	3,8
Speech	0	0	**100**	0
Gunshot	15,7	3,1	0	**80,2**

Table 2. *Two stage topology recognition rates (%) using the multi-domain feature set with respect to the sound classes related to non-human activities (average rate=100%, HMMs with 7 states and 16 Gaussian mixtures).*

Responded / Presented	Bird call	Rain	Wind
Bird call	**100**	0	0
Rain	0	**100**	0
Wind	0	0	**100**

*The HMM parameters with respect to the first stage of the system's topology are 4 states and 8 Gaussian components per state and the recognition rate is 87.7%.

reduce the dimensionality of the feature vector to 13 coefficients.

Audio pattern recognition algorithm

Sound recognition is based on the assumption that every sound source exhibits a consistent acoustic pattern which results in a specific way of distributing its energy on its frequency content. This unique pattern can be discovered and subsequently modeled by utilizing statistical pattern recognition algorithms. In this article we follow the HMM approach where each state is modeled by a Gaussian micture model (GMM) with a diagonal covariance matrix. HMMs have the ability to model the temporal evolution of sound events. This kind of pdf approximation is based on the assumption that the data follow a consistent temporal pattern which is expressed in terms of states.

The Baum-Welch algorithm was employed for training an HMM for each sound class. The algorithm breaks up the feature sequence into a predefined number of states and learns the associations between them. This results to a kxk transition matrix A whereas each one of its elements gives the probability of transition across different states. Thus, the element (i, j) is the probability of moving to state j at time $t+1$ given state i at time t. An HMM is defined by the following elements:

$$L = \{A, B, p_i\},$$

where B corresponds to the observation probabilities and p_i comprises the initial state distribution. We use left-right HMMs which means that there are no directed loops in the automation. Subsequently the models that have been already created are used for computing a degree of resemblance in the form of log-likelihood between each model and the unknown input signal. At this stage the Viterbi algorithm is employed. This type of score is computed with respect to all the constructed HMMs and the final decision is made by determining the maximum log-likelihood. Torch (available at http://www.torch.ch) implementation of GMM and HMM, written in C++ was used during the whole process. The maximum number of k-means iterations for initialization was 50 while both the EM and Baum-Welch algorithms had and upper limit of 25 iterations with a threshold of 0.001 between subsequent iterations.

3. FRAMEWORK CONSTRUCTION

In order to come up with the topology which provides the highest classification accuracy, we designed an experiment consisting of two phases. Initially we utilized the MFCCs and conducted a simple experiment with respect to two different classification topologies. The first one is consisted of one-stage and the second one uses an hierarchical schema which first discriminates the sound events which appear in the case of human presence vs the rest. Then another stage follows where the exact class of the novel sound event is predicted. The motivation behind experimenting with the two-stage approach lies in the fact that it limits the problem space as well as that the division is in line with the scope of this work as regards to identifying sound events related to human activities. One of the main burdens that sound recognition systems have to face is the decrease in their performance as the number of categories increases. Using the two-stage topology the largest number of categories that

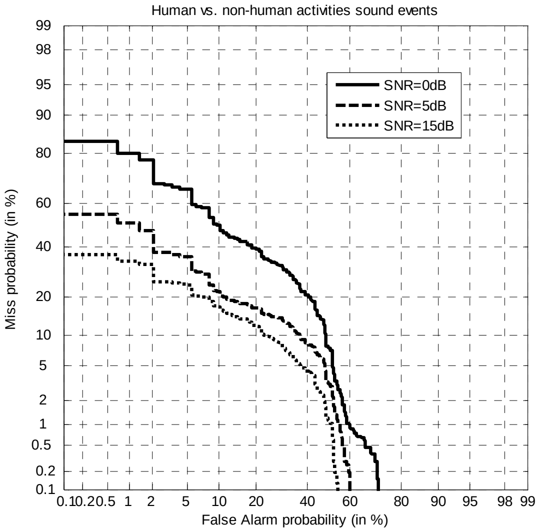

Figure 2: *DET curves for the detection of human activities sound events when merged with environmental noise. The SNR is equal to 0, 5 and 15dB while the corresponding EERs are 30.41%, 17.2% and 13.3% respectively.*

the system has to identify is four while in the case of the one-stage topology the corresponding number of classes is seven. The statistical models were optimized in terms of numbers of states as well as Gaussian components based on the highest recognition rate criterion. During every phase of our experimentations the parameters of the hidden Markov models were drawn from the following sets: a) number of states: {3, 4, 5, 6 and 7} and b) number of components: {2, 4, 8, 16, 32, 64, 128 and 256}. The models which provided the highest recognition accuracy were selected.

Feature set parameterization

Following the MPEG-7 standard recommendation, the low-level feature extraction window is 30 ms with 10 ms overlap, so the system is robust against possible misalignments. The sampled data are hamming windowed to smooth any discontinuities while the FFT size is 512. The number of the DCT coefficients is kept constant for all the feature sets and equal to 13.

Due to space limitation we provide results only for the concatenated feature vector which gave out the highest recognition accuracy. The achieved overall accuracy was equal to 81.5% while the recognition rate of the first stage was equal to 87.7%. The corresponding confusion matrices are demonstrated in Table 1 and Table 2. It should be noted that there is no overlap between train and test sets. The majority of the errors occurs due to the great variability (as it can be assessed by a human listener) among sound recordings of the same class. Furthermore, several sound clips are acoustically similar even though they belong to different categories. For example a large number of car movement events sound like motorcycle movements and vice versa. This fact is clearly observed by examining Table 1 which reflects upon the ability of the proposed system to discriminate sound events related to human activities. There we can see that the system confuses car with motorcycle sound events. The only class which has 100% recognition accuracy is the one dedicated to speech. Gunshot sound events are categorized correctly 80.2% of the time while the system tends to confuse them with motorcycle sound events. With respect to the sound events which are related to non-human activities we observe that the usage of the feature set based on the wavelet transform provided excellent recognition accuracy. Multidomain sets have the ability to capture various aspects of the information which is provided by the audio signal. Conclusively we could argue that after extensive experimentations, we managed to capture distinctive and characteristic information of the audio classes using a feature vector of rather low number of dimensions (39).

4. DETECTION OF HUMAN ACTIVITIES

Situations which include human activities in natural environments were artificially created by merging the corresponding audio signals. The merging of the audio signals was conducted at different energy ratios in order to observe the way that the system responds even at particularly difficult conditions. The Signal to Noise Ratios (SNR) which were used are 0, 5 and 15 dB. After merging each output is normalized by its maximum value in order to adjust the overall volume of the specific recording so that the strongest peak is at full level (gain normalization). Subsequently the respective sequence of feature coefficients is extracted and fed to the statistical models which provided the highest recognition accuracy during the previous experimental phase.

The detection experiment was conducted in the following manner: we merged every recording which is associated with human activities with a part of an environmental sound of equal size which is chosen randomly from the respective sound classes. This process is repeated 50 times for each recording so that all the recordings are merged with different and dissimilar parts of environmental sounds (for example for the motorcycle class we have 79x50=3950 different test samples). This ensures that the results are reliable and representative of the detection capabilities of the proposed system.

The Detection Error Tradeoff (DET) curves [5] which comprise an adapted version of Receiver Operating Characteristic (ROC) curves were used for evaluation. DET curves try to present the trade-off between missed detections and false alarms. The point where the average of the missed detection and false alarm rates is minimized is the optimal point, i.e. the one that should be used during the operation of the system. The specific average essentially is the cost function of a DET curve. When a large number of target events (in our case human activities) is available in combination with an almost equal amount of non-target events (environmental sounds), the performance of the system is demonstrated accurately. A convenient advantage of the DET against the ROC curves is the linearity which characterizes them. This property is highly desirable since it helps us towards choosing the system with the best performance quite easily.

During the first phase of the simulation experiments, target and non-target events were given as input to the two stage probabilistic framework and the log-likelihoods outputted by the human activities HMM were used for designing the respective DET plot. The specific plot is illustrated in Figure 2 for different SNR values and provides a picture of the detection capabilities of the system for all the sound events which are indicative of human activities (car, motorcycle, speech and gunshot) when merged with all the kinds of environmental noise (bird call, rain, wind). The Equal Error Rates (EERs) are equal to 30.41%, 17.2% and 13.3% while the SNR is 0, 5 and 15dB respectively. We observe that even under extremely noisy conditions (SNR=0dB), the proposed framework demonstrates quite good performance. As the SNR increases the detection rate is rapidly increased. By conducting listening tests with respect to the merged signals, it was derived that when the SNR is equal to 5dB, the real-world conditions are represented adequately. At the particular ratio, our system provided a relatively low EER which shows reliable detection of the sound events of interest. We conclude that the results analyzed are very encouraging and underline the importance of the selected statistical architecture in which features that capture diverse aspects of the audio structure were incorporated.

5. CONCLUSIONS

Detection of human activities like trespassing, hunting etc. in natural environments can very important towards their preservation. We analyzed a methodology for automatic acoustic detection of human presence in the specific type of environment. Our approach is based on a multi-domain description of the audio signals and a generative pdf estimation technique. An hierarchical topology was proposed and evaluated using recordings characterized by highly non-stationary environmental noise. The proposed approach is easily extendable to new audio classes as long as an adequate amount of training data is available.

Our future work includes elaborating on data captured on the Hymettus Mountain for the needs of the AMIBIO project[1]. Furthermore we intent to work on an adaptation module so as to provide a certain degree of flexibility and autonomy to the present surveillance framework.

6. ACKNOWLEDGEMENTS

This work was supported by the LIFE+ Program AMIBIO "Automatic Acoustic Monitoring and Inventorying of Biodiversity", Grant: LIFE08 NAT/GR/000539.

7. REFERENCES

[1] Ghiurcau, M.V., Rusu, C. and Bilcu, R.C., "Wildlife intruder detection using sounds captured by acoustic sensors," in ICASSP 2010, Dallas, Texas, March 2010.

[2] Ntalampiras, S., Potamitis, I. and Fakotakis, N. "An adaptive framework for acoustic monitoring of potential hazards," EURASIP Journal on Audio, Speech and Music Processing, vol. 2009, doi:10.1155/2009/594103.

[3] Quackenbush, S. and Lindsay, A., "Overview of MPEG-7 audio," IEEE Transactions on Circuits and Systems for Video Technology, 11(6): 725-729, June 2001.

[4] Ntalampiras, S., Potamitis, I., and Fakotakis, N., "Exploiting temporal feature integration for generalized sound recognition," EURASIP Journal on Advances in Signal Processing, vol. 2009, doi:10.1155/2009/807162.

[5] Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M., "The DET curve in assessment of detection task performance," in Eurospeech '97, Rhodos, Greece, Sept. 1997.

[1] http://www.amibio-project.eu/