Investeşte în oameni!
FONDUL SOCIAL EUROPEAN
Programul Operațional Sectorial Dezvoltarea Resurselor Umane 2007 – 2013
Axa prioritară: 1 „Educația şi formarea profesională în sprijinul creşterii economice şi dezvoltării societății bazate pe cunoaştere”
Domeniul major de intervenție: 1.5 „Programe doctorale si postdoctorale în sprijinul cercetării”
Titlul proiectului: Proiect de dezvoltare a studiilor de doctorat în tehnologii avansate- ”PRODOC”
Cod Contract: POSDRU 6/1.5/S/5
Beneficiar: Universitatea Tehnică din Cluj-Napoca

FACULTATEA DE ELECTRONICĂ, TELECOMUNICAȚII ŞI TEHNOLOGIA INFORMAȚIEI

Ing. **MARIUS VASILE GHIURCĂU**

REZUMAT TEZA DE DOCTORAT

SOUND CLASSIFICATION IN FORENSIC SCENARIOS

Conducător ştiințific
Prof.dr.ing. **CORNELIU RUSU**

1 Problem overview

All over the world, 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 reserves are the target of these illegal activities, but also the protected lakes or coastal regions (eg. 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 and it 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 the protected areas.

In the wildlife protection applications, standalone systems must be utilized. These systems will only send emergency messages to some remote observation station and must include all the classification and detection steps. This is because many areas under surveillance are remote and difficult to be visited, and also the intrusion of people for surveillance should be limited. As one can see, the implementation of a wildlife surveillance system requires the design of low complexity algorithms and the utilization of hardware with low power consumption.

Here we shall propose two solutions for detecting intruders in the wildlife regions. The first one is a low complexity solution, with low computation cost, which could be easily implemented on a simple controller. For the second solution, more complex approaches are utilized, increasing the accuracy of the system but at the same time increasing the complexity and the costs of the implementation.

Actually, this work is focused on 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 four sound classes: sounds originated from humans, birds, cars and animals.

A big problem in our task is represented by the fact that the sounds recorded in real environments using distant microphones are most of the times affected by various factors,

such as noise and reverberation. Because of this aspect, the accuracy of the sound recognition systems is strongly affected. Development of robust processing techniques for sound enhancement is a vital topic, which has been of a major interest for the scientific community for many years.

The second issue analyzed in this thesis is related to speaker recognition in an emotional environment. Usually the speaker recognition (identification) systems are trained using the voice of a certain person in a neutral (normal) state. But let us pretend that someone tries to use this kind of access system after a bad day, and his voice is rather bored or sad, than neutral. Or maybe someone noticed that he/she is being followed on his way home (or any other secured facility) and when arriving in front of the door and hurrying to get inside, its voice is rather attained by fear, scared, not really neutral. Will the identification system still identify and let him inside? Quite similar situations can be found in different forensic scenarios: a thief is robbing a bank, his face is covered, but when shouting at the bank employer, his voice is being recorded by the surveillance systems. The quality of the recording can be quite good, but even though, may one use this sample in order to compare it with the voice of a main suspect (suspects) that is being interrogated? His voice this time may be rather scared, or sad, but not angry or anxious, as in the first case.

This first chapter presents the work carried out and overviews the main motivations and problems encountered in the development of the present thesis. In the first section of the chapter the research topic is contextualized and the need for wildlife intruder detection systems is justified. Next, some examples of related work, for sound classification in wildlife regions, emotion recognition in speech and speaker recognition in an emotional environment are described. In the last section the main goals of the thesis are presented, followed by the organization of this document, with a short description of each chapter and the author's contributions.

2 Thesis Objectives

This thesis investigates and proposes solutions in two different fields of interest, both directly related to the audio classification domain. First part of the work is dedicated to the identification of intruders in wildlife regions. The work presented here represents only a first effort meant to open and explore this interesting and at the same time very necessary research topic. As a consequence, one of the main goals was to investigate the current state of the art and to establish a proper basis for future research and development in this area.

Closely related to the audio classification domain, this thesis is intended to provide an insight and some solutions to a topic of very recent research interest, namely speaker identification under an emotional environment.

3 Structure of the thesis

The research work undertaken in this thesis is structured as follows.

Chapter 1 introduces the topics, gives the motivations, some of the related work, the outline and the contributions of this PhD thesis.

The aim in *Chapter 2* is to provide the reader with the necessary background information on signal processing while familiarizing him/her with the audio signal and its relevant issues. The second part of the chapter discusses the most important details that concern speaker recognition and its applications.

Chapter 3 considers the audio classification algorithms used in this work. Detailed explanations for both the low complexity solutions (TESPAR: Time-Encoded Signal Processing And Recognition) and the more complex approaches (GMM: Gaussian Mixture Models and SVM: Support Vector Machines) are provided.

The first part of *Chapter 4* provides descriptions of the audio databases used in this study. The second part of the chapter focusses on the practical work undertaken, presenting in detail the experimental setup of the studies.

Chapter 5 highlights the results of the first study performed. This chapter includes several parts, each of them corresponding to a certain experiment. One by one, all the methods that were proposed for testing the accuracy of a potential wildlife intruder detection system are tested. In the end, a comparison of the results obtained in the experiments is made.

In *Chapter 6* is presented the result of the second study. The effect of the emotional state of a speaker upon the results of a text-independent speaker recognition application is analyzed. At the end of the chapter a short discussion of some possible solutions for increasing the accuracy of such a system in the emotional environment is presented.

Chapter 7 provides some general conclusions to the work by highlighting the major findings and contributions, along with a discussion of possible future research directions which arise from the research work undertaken in this thesis.

4 Author's contributions

Four databases with recordings from humans, cars (vehicles), birds and various animals were constructed. Both low complexity solutions and also standard sound classification algorithms were proposed for testing the accuracy of a potential intruder detection system. Also, it was proposed a modification of the standard TESPAR algorithm which improved the accuracy of the wildlife intruder detection system that was simulated.

It was studied the effect of the emotional state upon the variation of the fundamental frequency of a speaker and upon text-independent speaker recognition. The results showed

an important influence of the emotional state upon the accuracy of a speaker recognition system and also on the mean fundamental frequency. It was proposed a solution that increases the accuracy of a speaker recognition system in an emotional environment. The structure of the thesis is constructed to follow easily these results. The relevant contributions can be found in the following publications:

- In [Ghiurcau et al., 2010d] a solution for classifying sounds in wildlife regions is proposed. An implementation of a low complexity algorithm (TESPAR) was realized. TESPAR proved to be quite efficient in the classification of wildlife sounds and also fairly robust when noisy environments were simulated. The results were published at ICASSP'10.

- A modified version of the previous TESPAR algorithm is proposed in [Ghiurcau et al., 2010c]. The results show an improvement, in comparison to the previous case, and also a decrease in the amount of computations, which leads to energy savings in a hardware implementation. The work was presented at ISCAS'10.

- In [Ghiurcau et al., 2010b] a more complex solution that uses GMMs and MFCCs for detecting intruders in wildlife regions is proposed. A new database containing animal sounds is introduced. In this case, the accuracy of the system proposed is higher than the one of the modified TESPAR system, but at the cost of an increased complexity. The results were published at EUSIPCO'10.

- Another method for wildlife sound classification that uses SVMs and MFCCs is proposed in [Ghiurcau and Rusu, 2010]. The results are comparable to the ones obtained in [Ghiurcau et al., 2010b] and significantly better than the ones obtained in the first two studies. Again, as a compromise, the complexity of the system is increased. The results were presented at ISEEE'10.

- An overview of the audio based solutions for detecting intruders in wildlife regions in presented in [Ghiurcau et al., 2011c]. A comparison of all the proposed methods, pointing out the advantages and disadvantages of each of them is made. New experiments, involving all the databases are performed and the results are reported. In the end, a solution that uses both the low complexity and the increased complexity solutions is suggested.

- In [Ghiurcau et al., 2010a], a study of the effect of the emotional state upon the variation of the fundamental frequency of a speaker is performed. The results show an important influence of the emotional state on the fundamental frequency and also on the its standard deviation. This study was published in the Journal of Applied Computer Science and Mathematics (2010).

- The effect of the emotional state upon text-independent speaker identification is studied in [Ghiurcau et al., 2011b]. It is proven that, even though in the non-emotional situations the accuracy of the system is high (99%-100%), when emotions alter the human voice, the accuracy of the system drops significantly. In the end, a solution which increases the accuracy of the system is presented and some other future work possibilities are proposed. The results were published at ICASSP'11.

- Another method for testing the accuracy of speaker recognition systems under emotional environment was tested in [Ghiurcau et al., 2011a]. A comparison with the previous attempt is made. GMMs prove to be more suitable than SVMs when emotions alter the human voice. Still, the results are not fully satisfactory and a combined method of the GMMs and SVMs, or the use of other features is suggested. This study was presented at SPAMEC'11.

5 Conclusions

In the work presented here we have investigated and proposed solutions in two different fields of interest, both directly related to the audio classification domain. Firstly, we have evaluated the performances of a potential monitoring system that could be assimilated to an *acoustic eye*. The role of such a system is to determine possible intruders in various protected wildlife regions, such as natural reserves, protected lakes or coastal regions. These natural reserves with wildlife, flora, fauna or features of geological or other special interest, are spread and there is practically impossible a continuous surveillance. Even though these regions are protected by the law, they are the target of various illegalities, and preserving them is a serious issue for the authorities. Overall, systems of monitoring these wild regions and detecting intruders are very necessary, and would probably ensure a better preserving of these areas.

In all the experiments, closed set identification was performed. Both complex and low-complexity solutions were employed and the results are compared. The low-complexity approach proved to be fairly robust when various noise environments were simulated. As one may expect, the standard sound classification methods presented proved to be more robust than the low complexity solution. However, we are aware that such a complex system is hard to be implemented on a cheap controller and placed in a wildlife region. Even though, a possible combined solution that overcomes this difficulty will be tried.

For the future we want 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 could 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. One of the future goals is to increase our databases, with sounds that reproduce gun shot, thunder, chain saws, etc.

For the intruder verification system 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 cause a *system jam*.

In the second study we have we have evaluated the effect of the emotional state of a speaker when text-independent speaker identification is performed. Emotions are present in our everyday life and most of the time we cannot control them. Speaker recognition systems are very popular these days and they are used in various applications. Consequently there is a question that arises: how does the emotional state influence the accuracy of a speaker recognition system? In order to find the answer to this question several experiments were performed. In these experiments the spectral features used for speaker recognition were the Mel-frequency cepstral coefficients, while for training of the speaker models and testing the system the Gaussian Mixture Models and Support Vector Machines were used. The tests are performed on the Berlin emotional speech database which contains 10 different speakers recorded in different emotional situations: happy, angry, fear, bored, sad and neutral. As it was expected, SVM and GMM perform very well in text-independent speaker verification. When emotions alter the human voice, the performances of the speaker verification systems decrease significantly. If training of the system is done using utterances of the speakers in different emotional states, the correct verification rates increase up to approximately 98%. As a consequence, for the future it would be a good idea to train a speaker identification/verification system with utterances in different emotional states. Even though the increase is considerable in this case, it is still not sufficient. In order to be a viable solution, the rates should hold up to over 99%.

A possible solution could be using a 'two-step' identification system, in which firstly the emotion is identified, and furthermore we proceed to speaker identification using trained models of the particular identified emotion. For improvement of the existing system a combination of GMM and SVM should be tried. Using MFCC combined with other features, such as fundamental frequency, may improve the results, even though, as it was proven in the current study, fundamental frequency is a prosodic speech feature strongly dependent of the emotional states.

References

[Ghiurcau et al., 2010a] Ghiurcau, M. V., Lodin, A., and Rusu, C. (2010a). A study of the effect of emotional state upon the variation of the fundamental frequency of a speaker. *Journal of Applied Computer Science and Mathematics*, 4(7):79–82.

[Ghiurcau and Rusu, 2010] Ghiurcau, M. V. and Rusu, C. (2010). About classifying sounds in protected environments. In *Proceedings of The 3rd International Symposium on Electrical and Electronics Engineering (ISEEE 2010)*, pages 84–87, Galati, Romania.

[Ghiurcau et al., 2011a] Ghiurcau, M. V., Rusu, C., and Astola, J. (2011a). Speaker recognition in an emotional environment. In *Proceedings of Signal Processing and Applied Mathematics for Electronics and Communications (SPAMEC 2011)*, pages 81–84, Cluj-Napoca.

[Ghiurcau et al., 2011b] Ghiurcau, M. V., Rusu, C., and Astola, J. (2011b). A study of the effect of emotional state upon text-independent speaker identification. In *Proceedings of The 36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, pages 4944–4947, Prague, Czech Republic.

[Ghiurcau et al., 2010b] Ghiurcau, M. V., Rusu, C., Astola, J., and Bilcu, R. (2010b). Towards an application for detecting intruders in wildlife regions. In *Proceedings of The 2010 European Signal Processing Conference (EUSIPCO 2010)*, pages 1865–1868, Aalborg, Denmark.

[Ghiurcau et al., 2011c] Ghiurcau, M. V., Rusu, C., Astola, J., and Bilcu, R. (2011c). Audio based solutions for detecting intruders in wild areas. In *accepted for publication in Signal Processing*.

[Ghiurcau et al., 2010c] Ghiurcau, M. V., Rusu, C., and Bilcu, R. (2010c). A modified TESPAR algorithm for wildlife sound classification. In *Proceedings of The IEEE International Symposium on Circuits and Systems (ISCAS 2010)*, pages 2370–2373, Paris, France.

[Ghiurcau et al., 2010d] Ghiurcau, M. V., Rusu, C., and Bilcu, R. (2010d). Wildlife intruder detection using sounds captured by acoustic sensors. In *Proceedings of The 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, pages 297–300, Dallas, USA.