# ONE CLASS SUPPORT VECTOR MACHINES FOR AUDIO ABNORMAL EVENTS DETECTION

*Sébastien Lecomte[1,2], Régis Lengellé[2], Cédric Richard[3], François Capman[1]*

[1]Laboratoire Multi-Media Processing, Thales Communications, Colombes, France
[2] Institut Charles Delaunay - LM2S, UMR STMR, Université de Technologie de Troyes, France.
[3] Institut Universitaire de France, Laboratoire Fizeau, UMR, CNRS 6525, Observatoire de la Côte d'Azur, Université de Nice Sophia-Antipolis, France.

**ABSTRACT**

*This paper proposes an unsupervised method for real time detection of abnormal events in the context of audio surveillance. Based on training a One-Class Support Vector Machine (OC-SVM) to model the distribution of the normality (ambience), we propose to construct sets of decision functions. This modification allows controlling the trade-off between false-alarm and miss probabilities without modifying the trained OC-SVM that best capture the ambience boundaries, or its hyperparameters. Then we present an adaptive online scheme of temporal integration of the decision function output in order to increase performance and robustness. We also introduce a framework to generate databases based on real signals for the evaluation of audio surveillance systems. Finally, we present the performances obtained on the databases.*

***Index Terms***— One-Class SVM, unsupervised learning, detection, adaptive audio segmentation, audio surveillance.

## 1 INTRODUCTION

Third-generation surveillance systems (3GSS) [13] point out the interest in multimodal analysis of environments for public safety. This includes the use of audio in complement of video. The increasing demand for surveillance harms operators watchfulness as they are supposed to supervise dozens of screens and other sensors. In this context, audio analysis might provide automatic awareness in order to focus operators attention on real risky situations. The classical framework for doing so involves two main steps: 1- detection of abnormal situations, 2-recognition of detected events. In this paper, we focus on improving the detection stage.

Most of the audio surveillance systems proposed in the literature are supervised; they need more than the training signal itself to be trained. Minimum information requirement is a challenge for building more efficient and intelligent automated surveillance systems [18]. Thus, we only consider here non supervised systems.

Classical approaches consist of setting up detectors dedicated to few identified abnormal events [17], or strongly related to the *a priori* ambience [10]. These are not suitable in the context of surveillance as: 1- we have no prior information relative to events to detect; 2- in noisy environments (e.g. public transports, or urban complexes), the ambience is a non-stationary continuum that may include normal sound events. We propose to use non-supervised approaches to learn ambience patterns. Literature offers methods to estimate the distribution of normality; Gaussian Mixtures Models (GMM) and One Class Support Vector Machine (OC-SVM) being the most popular. Because of the nature of the optimized criterion, OC-SVM [14] presents better generalization results. Furthermore, we do not have any prior on the ambience data distribution in the acoustic space and OC-SVM algorithms are able to model arbitrarily shaped sets. For these reasons, we propose to capture the distribution of the ambience with OC-SVM and we consider every rejected point as a detected event to be classified. We also slightly modify OC-SVM to deal with the trade-off between detection and false alarm probabilities.

In order to improve the detection scores of our detectors, we temporally integrate the detector output. As events are of variable length, we show that an optimal integration requires segmentation information. Most of the proposed audio segmentation algorithms in the literature are based on information criterion such as Bayesian Information Criterion (BIC), [3]. Unfortunately, it is difficult to design a segmentation module with good generalization capabilities when considering various types of audio signals. We propose here an automatic on-line segmentation module based on multi-level segmentation, as already suggested in the area of speech recognition [5] [6].

This paper begins with the presentation of One-Class SVMs and the construction of the family of decision functions. Section 3 gives an overview of the proposed segmentation for temporal-integration of the decision function. Then, section 4 introduces the generation of an evaluation database from real signals, adding amplitude controlled abnormal events to ambience signals. We present experimental results in section 5. Finally, we conclude with some perspectives.

## 2 OC-SVM BASED DETECTOR FOR AUDIO SIGNALS

### 2.1 OC-SVM basic elements

Let $\{x_1 \dots x_l\}, x_i \in X \subset \mathbb{R}^d$ be the training set, where $l \in \mathbb{N}$ is the number of observations that belong to a single class, ambience signals in our application. OC-SVM [14] aims to define the boundary of $\Gamma$, the minimum volume region enclosing $(1-\nu)l$ observations. Hyperparameter $\nu$, in $[0;1]$, controls the fraction of observations that are allowed to be out of $\Gamma$ (outliers). Let $f_X: \mathbb{R}^d \to \mathbb{R}$ be a decision function such that:

$$f_X(x) \geq 0, \text{if } x \in \Gamma$$
$$f_X(x) < 0, \text{otherwise}$$

Within the context of SVM, the space of possible functions $f_X(x)$ is reduced to a Reproducing Kernel Hilbert Space (RKHS) with kernel $\kappa: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. This kernel induces the so-called feature space $H$ via the mapping $\phi: \mathbb{R}^d \to H$. Let $\langle .,. \rangle_H$ be a dot product in $H$. We consider here the Gaussian kernel:

$$\kappa(x,x') = \langle \phi(x), \phi(x') \rangle_H = \exp(-\|x-x'\|^2/2\sigma^2)$$

Since in this case $\kappa(x,x) = 1$, all the data are mapped onto the unit-radius hypersphere centered at the origin of $H$.

Training an OC-SVM consists of defining the separation hyperplane $W = \{h \in H \; s.t. \langle h, w \rangle_H - b = 0\}$ such that the margin $b/\|w\|_H$ is maximum (see Figure 1). Parameters $w$ and $b$ result from the optimization problem [14]:

$$\min_{w,\xi,b} \frac{1}{2}\|w\|_H^2 - b + \frac{1}{\nu l}\sum_{i=1}^{l} \xi_i \text{ subject to } \begin{cases} \langle w, \phi(x_i) \rangle_H \geq b - \xi_i \\ \xi_i \geq 0, i = 1 \dots l \end{cases}$$

where $\xi_i$ are the slack variables representing the loss associated with $x_i$ (non null $\xi_i$ define the outliers). The Lagrangian multipliers $\boldsymbol{\alpha}$ associated with this problem fully determine $w$ and $b$. We finally obtain the decision function $f_X(x) = \sum_i \alpha_i \kappa(x_i, x) - b$ and $\{x_j / \alpha_j \neq 0\}$ defines the support vectors set.

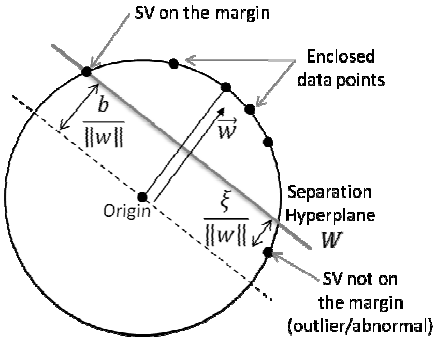


*Figure 1- Principle of One-Class SVM*

### 2.2 Constructing OC-SVM based decision function set

In standard OC-SVM, an observation is declared as belonging to the ambience class ($H_0$) if $f_X(x) \geq 0$. Once this decision rule is obtained, we propose to define a family of decision rules by modifying the threshold:

$$\begin{cases} \text{if } f_X(x) \geq \lambda, x \text{ is declared as ambience } (D_0) \\ \text{if } f_X(x) < \lambda, x \text{ is declared as abnormal event } (D_1) \end{cases}$$

This formulation allows control of the trade-off between miss and false-alarm probabilities $P(D_0|H_1)$ and $P(D_1|H_0)$ respectively. The introduced threshold $\lambda$ might be determined experimentally by operational requirements. $\lambda$ controls a translation of the separating hyperplane $W$, in the feature space $H$. Then, the resulting boundary of $\Gamma$, in the representation space, is the contour of the decision function $f_X(x)$ given $\lambda$.

Choosing $\nu$ is a challenging problem as it directly drives the fraction of training data lying inside the domain. This is conditioned by the application and operational requirements, in terms of detection rates or miss/false-alarm probabilities. In fact, for small values of $\nu$, $\Gamma$ can be estimated on regions of the input space where the density of training data is very low; increasing the variance of the estimation of $\Gamma$. Conversely, high values of $\nu$ can result in a high bias of the estimate of $\Gamma$ which could poorly represent the ambience data distribution. In our approach, we select an optimal value of $\nu$ that is only driven by the quality of the input signal, estimating the number of potential outliers, then we try to raise the performance to operational requirements using a suitable threshold $\lambda$.

Figure 2 illustrates, based on an experimentation for abnormal audio event detection, the decision function sets obtained with different values of $\nu$. Any set of decision functions corresponds to a curve, showing the balance between miss and false-alarm probabilities. For each curve, $\lambda$ varies from $\min_{\forall x \in T} f_X(x)$ to $\max_{\forall x \in T} f_X(x)$ where $T \subset \mathbb{R}^d$ is the evaluation set. On each curve, a square symbol locates the default OC-SVM decision function performance ($\lambda = 0$). Preliminary results show that, for audio surveillance signals, a good choice for $\nu$ can lead to better performance as $\lambda$ varies, compared to varying only $\nu$.

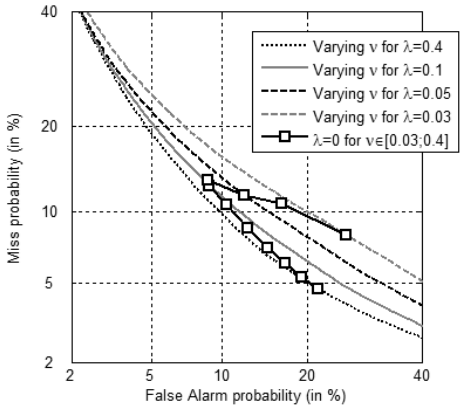


*Figure 2- Influence of $\nu$ and $\lambda$ on detection false-alarm / miss probability trade-off. Squares represent the decision function obtained with a standard OC-SVM ($\lambda = 0$).*

### 2.3 Temporal integration of decision function

The audio stream is first processed frame-by-frame. Each frame, composed of successive audio samples of a few tenths of milliseconds, is processed to extract a vector of "acoustic" descriptors. Then we compute the decision function from these vector. This approach does not correspond to an operational need and ignores temporal aspects which are now dealt with. We first propose the following median-filtered decision function:

$$\begin{cases} \text{if } \mathcal{M}_M\big(f_X(x_k)\big) \geq \lambda, x \text{ is declared as ambience } (D_0) \\ \text{if } \mathcal{M}_M\big(f_X(x_k)\big) < \lambda, x \text{ is declared as abnormal event } (D_1) \end{cases}$$

where $x_k$ is the k-th feature vector, $\mathcal{M}_M(u_k)$ the operator returning the median value of the series $\{u_{k-M+1}, \dots, u_k\}$ and $M$, the filter order. This filter provides good performance when event duration corresponds to the duration of integration, fixed by $M$. This observation is incoherent with the fact that events are of variable length. Therefore, we now present an adaptive integration scheme.

## 3 AUTOMATIC ONLINE SEGMENTATION OF AUDIO

The segmentation process that we use is a multi-level online algorithm (see Figure 3). First, in order to dispose of a long term representation of the signal, we feed a buffer with frame-by-frame extracted feature vectors. With no *a priori* knowledge of the signals, it seems reasonable to use standard spectral representation as features. Therefore, we use output energies of a Fourier-based linear filter bank. No assumption is possible regarding the duration of audio segments but the system should react as fast as possible. Few seconds can be considered as an operational requirement for surveillance applications. This gives the buffer size.

Once the buffer is full, we start the adaptive segmentation process. Based on a pre-defined similarity criterion, the closest successive pairs of segments are iteratively merged until only one segment remains (bottom-up hierarchical-clustering-based merging process). We use the Euclidean distance between mean vectors of

segments. In the resulting structure (that we can represent as a dendrogram), we look for the optimal segmentation level. A typical criterion consists of applying some pre-defined threshold to the distance of the closest merged segments. In our implementation, we computed the intra-segment correlation coefficient (correlation between merged segments), and chose the segmentation level which provides correlation coefficients above a given threshold for all segments. The segmentation at this level was kept except for the last segment that contains the first frames of the next buffer segment.

Then, the decision statistic is integrated over each homogeneous segment. For every segment, we now define the final decision function:

$$\begin{cases} \text{if } \langle f_X(x_k)\rangle_{S_j} \geq \lambda, x \text{ is declared as ambience } (D_0) \\ \text{if } \langle f_X(x_k)\rangle_{S_j} < \lambda, x \text{ is declared as abnormal event } (D_1) \end{cases}$$

where $\langle u_k \rangle_{S_j}$ is the operator returning the average value of the series $\{u_k / \forall k \in S_j\}$ and $S_j$ the set of frame indices belonging to segment $j$.

In all our experiments (as in Figure 4), buffer length was 2000ms (200 frames), fed with linear 24-filterbank output energies (features for segmentation can be different from those used for detection). Segments were constrained to contain from 8 to 100 frames and the quality measure threshold is set to 0.98.

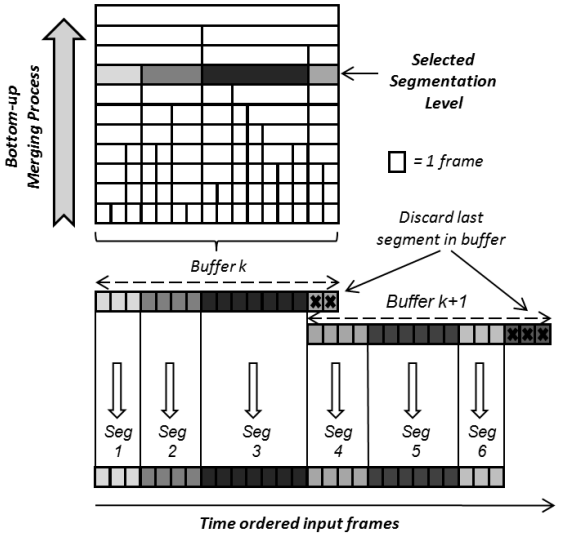


*Figure 3- Multi-Level Segmentation illustration.*

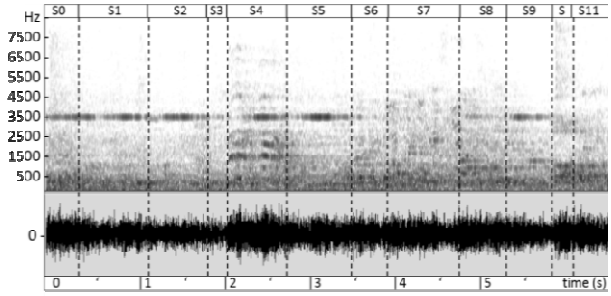


*Figure 4- Example of segmentation results on a recording from a Rome subway station: waveform (bottom), spectrogram (middle), segmentation result (top); vertical dashed lines separate segments.*

## 4 DATABASE FOR AUDIO SURVEILLANCE

We have developed a framework to combine abnormal events with recorded real ambience signals from a place under surveillance (see Figure 6). Our approach has several advantages: control of SNR, perfect knowledge of events position, fast and flexible generation of a large amount of test signals. In our application, the SNR is defined globally. Practically, this means that the SNR is computed globally from the average energy over the whole ambience signal. The event is then duplicated and inserted with constant amplitude at different positions into the ambience signal. This approach allows building signals representative of the variability of real ambience signals.

Literature is extensive concerning the measure of noise level [4][8][1]. Figure 5 indicates empirical measures of SNR over surveillance signals using different standardized weighting functions [7][9]. These results show that using ITU-R468 results in higher values. This indicates that this weighting is more adapted in evaluating the SNR of events into surveillance signals. Therefore, we use this weighting for targeting SNRs when generating the database.

| Weighting | Flat | A-type | C-type | ITU-R468 |
|---|---|---|---|---|
| SNR | 10,77 dB | 12,61 dB | 10,5 dB | 15,09 dB |

*Figure 5- Experimental SNRs using different weighting functions*

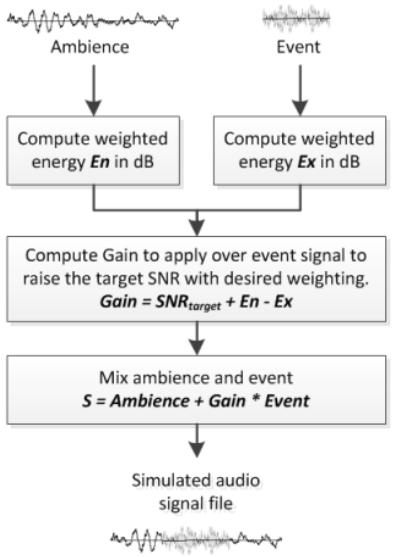


*Figure 6- Simulation flowchart of surveillance audio signals to evaluate abnormal event detection system*

## 5 EVALUATION

We generated a database using 18 ambience signals of 10 minutes each. Events are 1-second sounds extracted from commercial databases [15]. We considered 96 events grouped into 27 categories. Ambiences were recordings from a subway station in Rome [2]. Training set was composed of 6 ambience signals (1 hour). Each event was introduced at 50 different positions into each of the 12 remaining ambience signals (used as a test set). We selected 5 SNRs varying from 5 to 25dB. This process resulted in 5,760 audio files with an abnormal event occurring every 12 seconds. This corresponds to 288,000 events and 960 hours of

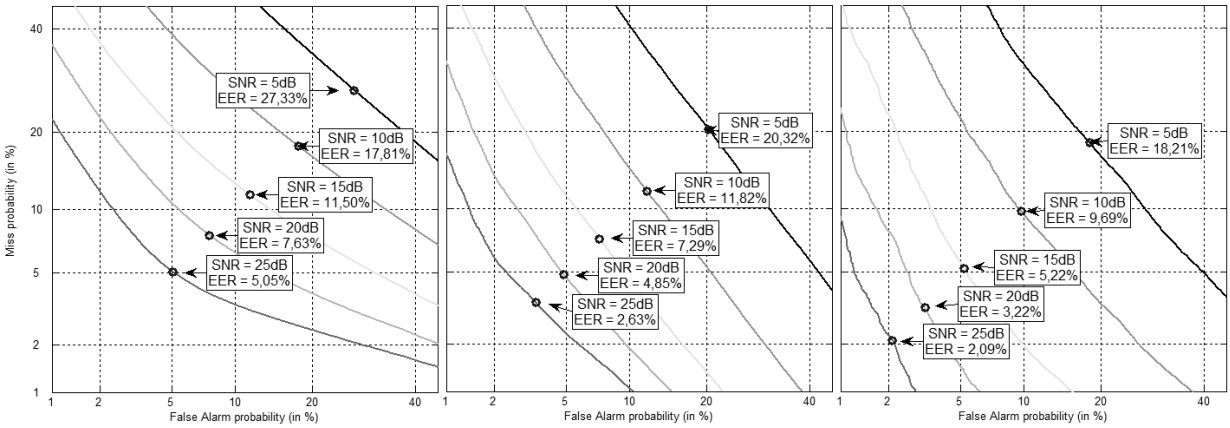


*Figure 7 - Detection Error Trade-off curves at different SNR without segmentation (left), with the proposed 1-second median filter on decision output (middle), and with application of our adaptive multi-level segmentation process (right).*

audio for evaluation. All signals were analyzed frame by frame in an online framework. Frames were 20ms long, Hamming windowed, with a 50% overlap. A 32-dimensional feature vector composed of the output energies of a linear filterbank was computed for each frame. Then we performed the online multi-level segmentation presented in section 3. OC-SVM hyperparameters were set to $\sigma = 10$ and $\nu = 10^{-3}$, resulting in 1548 support vectors from 360,033 learning data. Evaluation consisted of applying our decision function 1- frame by frame, 2- after median filtering over 1 second and 3- using adaptive segmentation. Obtained results are presented in Figure 7 [10]. Each curve corresponds to a given SNR and illustrates the trade-off between miss and false-alarm probabilities.

## 6 CONCLUSION

In this paper, we introduced a method for online detection of abnormal events in audio signals. Based on One-Class Support Vector Machines, our approach consists of a slight modification on the decision function. We also discussed the need for temporal integration of detection functions and developed a multi-level segmentation algorithm. Then we described a process for elaborating non standards audio test databases applied to surveillance. Finally, we presented experiments that illustrated the efficiency of our unsupervised framework. Future work will focus on studying relations between hyperparameters and the signal shapes, and criteria for measuring their quality.

## 7 ACKNOWLEDGEMENT

This work is supported by French National Association of Research and Technology (ANRT): CIFRE n° 970/2009; and the research leading to these results has received funding from the European Community's Seventh Framework Program FP7/2007-2013 – Challenge 2- Cognitive Systems, Interaction, Robotics – under grant agreement n° 248907-VANAHEIM. The authors would like to thank Mireille Tohmé [14] for her valuable work on $Fast - OC_2$ solver and Bertrand Ravera for his support.

## 8 REFERENCES

[1] British Broadcasting Corporation, "The Assessment of Noise in Audio-Frequency Circuits", *Engineering Division Research Report*, EL17, 1968.

[2] "CARETAKER Project: Content Analysis REtrieval Technologies to Apply Knowledge Extraction to massive Recording", 2006-2008, FP6 IST 4-027231.

[3] Cettolo, M., Vescovi, M., Rizzi, R., "Evaluation of BIC-based Algorithms for Audio Segmentation", *Computer Speech & Language*, pp.147-170, vol.19, issue 2, 2005.

[4] Fletcher, H, and Munson, W.A., "Loudness Its Definition, Measurement and Calculation", *J. of Acoustic Soc. of America*, pp.82-108, vol.V, 1933.

[5] Glass, J.R. and Zue, V. W., "Multi-Level Acoustic Segmentation of Continuous Speech", *Intl. Conf. on Acoustics, Speech, and Signal Processing*, IEEE, pp.429-432 vol.1, 1988.

[6] Husson, J.-L., Laprie, Y., "A new Search Algorithm in Segmentation Lattices of Speech Signals", *Proceedings of the 4th International Conf. on Spoken Language Processing*, IEEE, pp.2099-2102, vol.4, 1996.

[7] International Electrotechnical Commission, "Sound Level Meters - Part 2: Pattern Evaluation Tests", *29-Electroacoustics*, IEC-61672-2,1.0, 2003.

[8] International Organization for Standardization, "Acoustics – Normal Equal-Loudness-Level Contours", *ISO Standards*, ISO226, 2003.

[9] International Telecommunication Union, "Measurement of Audio-Frequency Noise Voltage Level in Sound Broadcasting", *Recommendation, Broadcasting Service*, ITU-R BS.468-4 1986.

[10] Istrate, D., Castelli, E., Vacher M., Besacier, L., and Serignat, J.-F., "Information Extraction From Sound for Medical Telemonitoring", *Trans. on Information Tech in Biomedicine*, IEEE, pp.264-274, vol.10, no.2, 2006.

[11] National Institute of Standards and Technology, "DET-Curve Plotting Software", *Information Technology Laboratory*, DETware v.2.1, [available online: http://www.itl.nist.gov/iad/mig/tools/]

[12] Rabaoui, A., Davy, M., Rossignol, S., Lachiri, Z. and Ellouze, N., "Improved One-Class SVM Classifier for Sounds Classification", *Conf. on Advanced Video and Signal Based Surveillance*, IEEE, pp.117-122, 2007.

[13] Räty, T.D., "Survey on Contemporary Remote Surveillance Systems for Public Safety", *Transactions on Systems, Man, and Cybernetics – Part. C: Applications and Reviews*, IEEE, pp.493-515, vol.40, no.5, 2010.

[14] Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola A. and Williamson, R.C., "Estimating the Support of a High-Dimensional Distribution", *Com. by Vladimir Vapnik – Neural Computation*, pp.1443-1471, vol.13, 2001.

[15] Sound Ideas, "The Series 6000 "The General" Sound Effects Library".

[16] Tohmé, M. and Lengellé, R., "Maximum Margin One Class Support Vector Machines for Multiclass Problems", *Submitted to Pattern Recognition Letters*, 2011.

[17] Valenzise, G., Gerosa, L., Tagliasacchi,M., Antonacci,F. and Sarti, A., "Scream and Gunshot Detection and Localization for Audio-Surveillance", *Conf. on Adv. Video and Signal Based Surveillance*, IEEE, pp.21-26, 2007.

[18] Valera, M. and Velastin S.A., "Intelligent Distributed Surveillance Systems: a review", *IEEE Proceedings - Vision Image Signal Processing*, pp.192-204, vol.152, no.2, 2005.