Abnormal Events Detection using Unsupervised One-Class SVM – *Application to Audio Surveillance and Evaluation* –

Sébastien Lecomte, Régis Lengellé
Institut Charles Delaunay - LM2S
UMR STMR Universit de Technologie de Troyes
12, rue Marie Curie - 10000 Troyes, FR
first.last@utt.fr

Cédric Richard
Institut Univ. de France
Laboratoire Fizeau, UMR, CNRS 6525
Observatoire de la Cte dAzur
Univ. Nice Sophia-Antipolis, FR
cedric.richard@unice.fr

Francois Capman, Bertrand Ravera
Thales Communications
146, bd de Valmy - 92700 Colombes, FR
first.last@fr.thalesgroup.com

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 generated database.

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. We also present and justify an innovative way to generate databases of audio signals for evaluation of audio surveillance systems.

Most of the audio surveillance systems proposed in the literature are supervised; they need more than the training signal itself to be trained whereas minimum information requirement is a challenge for building more efficient and intelligent automated surveillance systems [18]. Classical approaches are supervised and 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 strong prior information relative to abnormal events; 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 [12, 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 distribu-

This work is supported by French National Association of Research and Technology (ANRT): CIFRE n 970/2009. The research leading to these results has received funding from the European Communitys Seventh Framework Program FP7/2007-2013 Challenge 2- Cognitive Systems, Interaction, Robotics under grant agreement n 248907-VANAHEIM.

978-1-4577-0845-9/11/$26.00

tion 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 miss-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 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 introduces the temporal integration of the decision function and our automatic online segmentation process. Then, section 4 gives an overview of our scheme to generate evaluation databases, adding amplitude controlled abnormal events to real 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 $\{x_1 \dots x_l\}$, $x_i \in X \in \mathbb{R}^d$ 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 Γ, the minimum volume region enclosing $(1-\nu)l$ observations. Hyperparameter ν, in $[0;1]$, controls the fraction of observations that are allowed to be out of Γ (outliers). Let $f_X : \mathbb{R}^d \rightarrow \mathbb{R}$ be a decision function such that:

$$\begin{cases} f_X(x) \geq 0 & \text{if } x \in \Gamma \\ f_X(x) < 0 & \text{otherwise} \end{cases} \quad (1)$$

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

$$\kappa(x, x') = \langle \phi(x), \phi(x') \rangle_H = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2} \right) \quad (2)$$

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 $W = \{h \in H \text{ s.t. } \langle h, w \rangle_H - b = 0\}$ such that the margin $\frac{b}{\|w\|_H}$ is maximum (see figure 1). Parameters w and b are the solutions of the optimization problem [14, 16]:

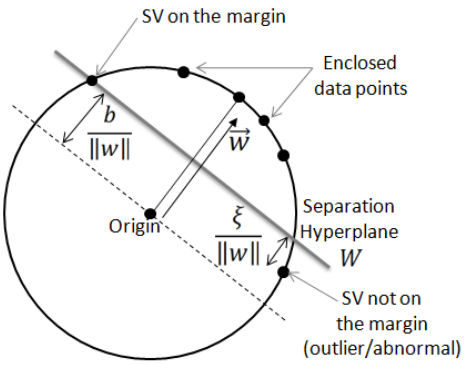

Figure 1. Principle of One-Class SVM

$$\min_{w,\xi,b} \frac{1}{2} \|w\|_H^2 - b + \frac{1}{\nu l} \sum_{i=1}^{l} \xi_i$$
$$\text{s.t.} \begin{cases} \langle w, \phi(x_i) \rangle_H \geq b - \xi_i \\ \xi_i \leq 0, i = 1 \dots l \end{cases} \quad (3)$$

where ξ_i are the slack variables representing the loss associated with x_i (non null ξ_i define the outliers). The Lagrangian multipliers α 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 $\{x_j / a_j \neq 0\}$ defines the support vectors set.

2.2. Constructing OC-SVM based Decision Function Set

Let H_0 be the hypothesis that a data x "belongs to the ambience class". In standard OC-SVM, if $f_X(x) \geq 0$ then x is declared as belonging to H_0 (D_0 decision). Considering hypothesis H_1 which is x "is an abnormal event" (D_1 decision), we now propose to define a family of decision rules introducing a threshold λ:

$$\begin{cases} \text{if } f_X(x) \geq \lambda, \text{ then } D_0\text{: ambience} \\ \text{if } f_X(x) < \lambda, \text{ then } D_1\text{: abnormal event} \end{cases} \quad (4)$$

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 λ might be determined experimentally by operational requirements. λ controls a translation of the separating hyperplane W, in the feature space H. Then, the resulting boundary of Γ, in the representation space, is the contour of the decision function $f_X(x)$ given λ.

Choosing ν 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 ν, Γ can be estimated

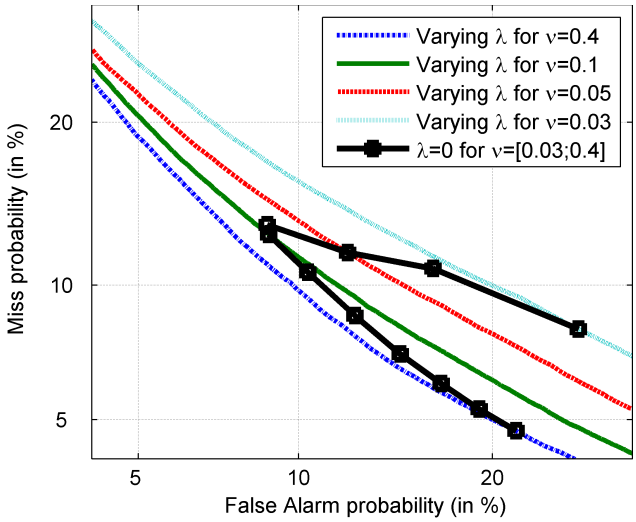

Figure 2. Influence of ν and λ on false-alarm / miss probability trade-off. Squares represent the decision function obtained with a standard OC-SVM ($\lambda = 0$).

on regions of the input space where the density of training data is very low; increasing the variance of the estimation of Γ. Conversely, high values of ν can result in a high bias of the estimate of Γ which could poorly represent the ambience data distribution. In our approach, we select an optimal value of ν 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 λ.

Figure 2 illustrates, based on an experimentation for abnormal audio event detection over a small database, the decision function sets obtained with different values of ν. Any set of decision functions corresponds to a curve, showing the balance between miss and false-alarm probabilities. For each curve, λ varies from $\min_{\forall x \in T} f_X(x)$ to $\min_{\forall x \in T} f_X(x)$ 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 ν can lead to better performance as λ varies, compared to varying only ν. In figure 2, square symbols represent the performances (in the sense of a trade-off between miss and false-alarm probabilities) of canonical uses of One-Class SVM (*i.e.* $\lambda = 0$) when $\nu \in \{0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.08, 0.05, 0.04, 0.03\}$. As one might expect, decreasing ν implies a reduction of the false-alarm probability; we also note that $P(D_1|H_0) \neq \nu$ due to the slight variation of ambiance between train and evaluation signals. When ν is too small, it is not possible for the OC-SVM algorithm to find a good representation of the data distribution and performances fall.

3. Temporal Integration

3.1. Principle

The audio stream is first processed frame-by-frame. Each frame, composed of successive audio samples of a few tenths of milliseconds (20 to 100 ms), 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.

First, we propose the following median-filtered decision function:

$$\begin{cases} \text{if } \mathcal{M}_M\left(f_X(x)\right) \geq \lambda, \text{ then } D_0\text{: ambience} \\ \text{if } \mathcal{M}_M\left(f_X(x)\right) < \lambda, \text{ then } D_1\text{: abnormal event} \end{cases} \quad (5)$$

where x_k is the kth feature vector, $\mathcal{M}_M(u_k)$ the operator returning the median value of the series $(u_{k-M+1}, \ldots, u_k)$ 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.

3.2. 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 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 is 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

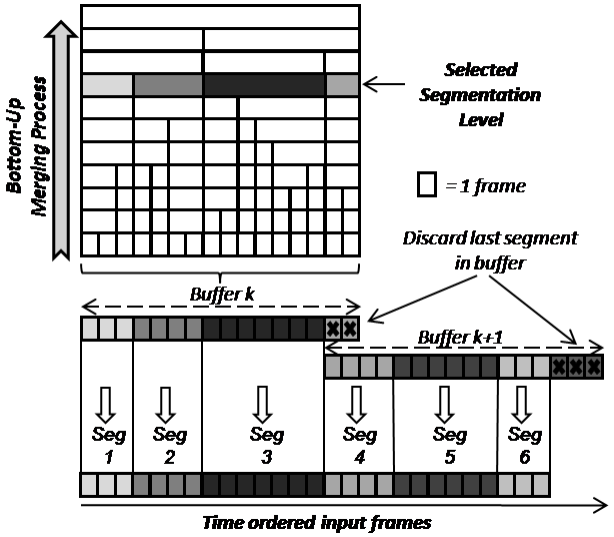

Figure 3. Multi-Level Segmentation illustration.

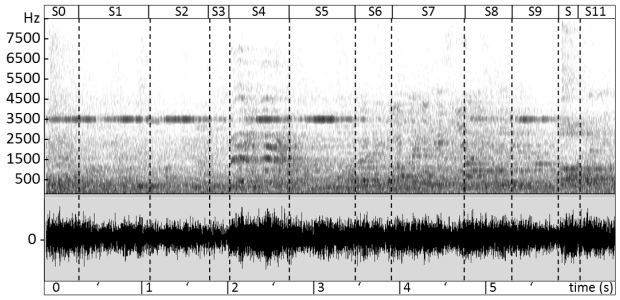

Figure 4. Example of segmentation results on a recording from a Rome subway station [2] waveform (bottom), spectrogram (middle), segmentation (top); vertical lines separate segments.

final decision function:

$$\begin{array}{l} \text{if} \quad f_X(x)_{\ S_j} \qquad \text{, then } D_0\text{: ambience} \\ \text{if} \quad f_X(x)_{\ S_j} < \quad \text{, then } D_1\text{: abnormal event} \end{array} \tag{6}$$

where $u_{k\ S_j}$ is the operator returning the average value of the series $u_k \quad k \quad S_j$ 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 intra-segment cross-correlation threshold is set to 0 98.

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 5). Our approach has several advantages: control of Signal to Noise Ratio (SNR), perfect knowledge of events position, fast and flexible generation of a large amount of test signals. Literature is extensive concerning weighted measures of noise level and SNR [1, 4, 9]. Table 1 indicates empirical measures of SNR over surveillance signals using different standardized weighting functions [7, 8]. These results show that using ITU-R468 [7] 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 targetting SNRs when generating the database.

Weighting	Flat	A-type	C-type	ITU-R468
SNR	10.77dB	12.61dB	10.5dB	15.09dB

Table 1. Experimental SNRs using different weighting functions

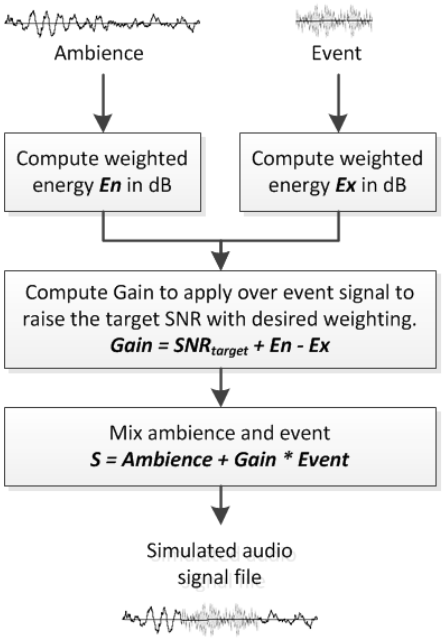

Figure 5. Simulation flowchart of surveillance audio signals to evaluate abnormal event detection system.

In our application, the SNR is defined globally. Practically, this means that a gain is computed from the average energy over the whole ambience and event signals to raise a global mean SNR. The event is then duplicated and inserted with constant amplitude at different positions into the ambience signal. The local SNR may varry from the targetted one for each inserted abnormal event. This approach allows building signals representative of the variability of real ambience signals and is coherrent with most of real life sounds (*e.g.* doors would not make louder noise if the ambiance noise is highly present).

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 (see figure 9). 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 6 SNRs varying from 5 to 30dB. This process resulted in $6,912$ audio files with an abnormal event occurring every 12 seconds. This corresponds to $345,600$ events and 1152 hours of 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. OC-SVM hyperparameters were set to $\sigma = 10$ and $\nu = 10^{-3}$, resulting in 1473 support vectors from $360,033$ learning data.

Figure 6 shows the distribution of local variations of SNRs using recorded ambiance signals [2]. Figure 7 presents three Detection Error Trade-off curves (DETcurves [11]) for a target SNR of 15dB. The DET curves are obtained considering all events included in ambience signals (*i.e.* all local SNRs), then considering only the obtained events resulting in locals SNRs lying in an interval of width 5dB or 2dB and centered at the targetted SNR. The curve shapes are the same, and they just slightly differ; this behavior is the same at each targetted SNR. Moreover, performance are always better when considering a reduced range than using all SNRs. This last consideration means that performing detection evaluation using all local SNRs results in a pessimistic evaluation of detection performances at a target global SNR, on this generated database.

Evaluation of the detection capabilities of our system consisted of applying our decision function 1- frame by frame (4), 2- after median filtering over 1 second (5) and 3- using the proposed adaptive segmentation process (6). Obtained results targetting global SNRs from 5dB to 25dB are presented in figure 8. Each curve corresponds to a given SNR and illustrates the trade-off between miss and false-alarm probabilities when λ varies from $\min_{\forall x \in T} f_X(x)$ to $\min_{\forall x \in T} f_X(x)$ where $T \subset \mathbb{R}^d$ is the evaluation set.

Figure 9 details in term of Equal Error Rate (EER), *i.e.* $P(D_0|H_1) = P(D_1|H_0)$, the results obtained for each 27 considered type of event at different global targetted SNRs from 10dB to 30dB. For each event type, when the SNR increases, the performances also increase until some threshold is raised. This is due to the fact that at some SNRs abnormal events might be very close to events learned from the ambience: *e.g.* the noise from train brakes when arriving at the station is spectrally close to loud cheering abnormal events, or similarly, loud fight abnormal events (impulsive events) cause signal saturation like the closing doors of the trolleys.

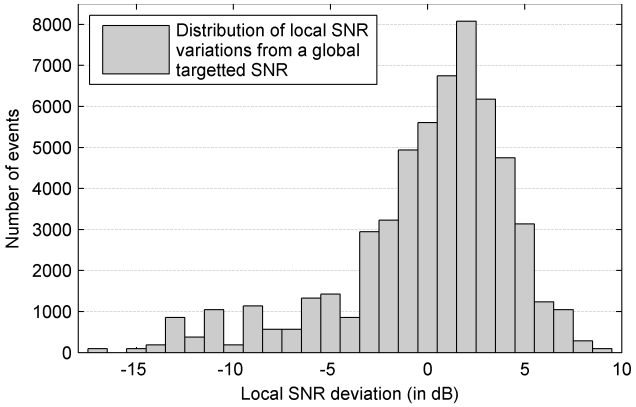

Figure 6. Distribution of local SNR variations while adding an event into an ambience signal using our method. The average variation is 0.248dB and standard deviation is 4.756dB.

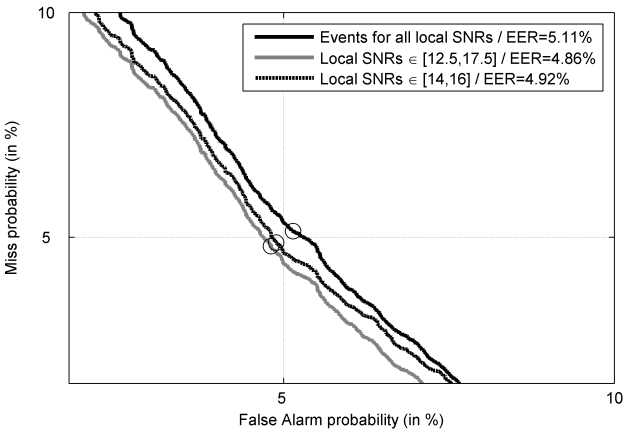

Figure 7. DETcurves when targetting a SNR of 15dB considering 1- all events (*i.e.* all values of local SNRs), 2- events whose local SNR are in a 5dB window centered around the target SNR and 3- idem in a 2dB window.

6. Conclusion

In this paper, we presented a complete scheme for online detection of audio abnormal events in surveillance applications; taking into account integration and use-case constraints. This framework integrates an unsupervised learning step, then an online evaluation step. Based on One-Class Support Vector Machines, our approach allows the construction of decision function sets for a fixed choice of hyperparameters. This also allows modifying the detector characteristics without performing a new train. We also discussed the need for temporal integration of detection functions and presented a multi-level segmentation algorithm

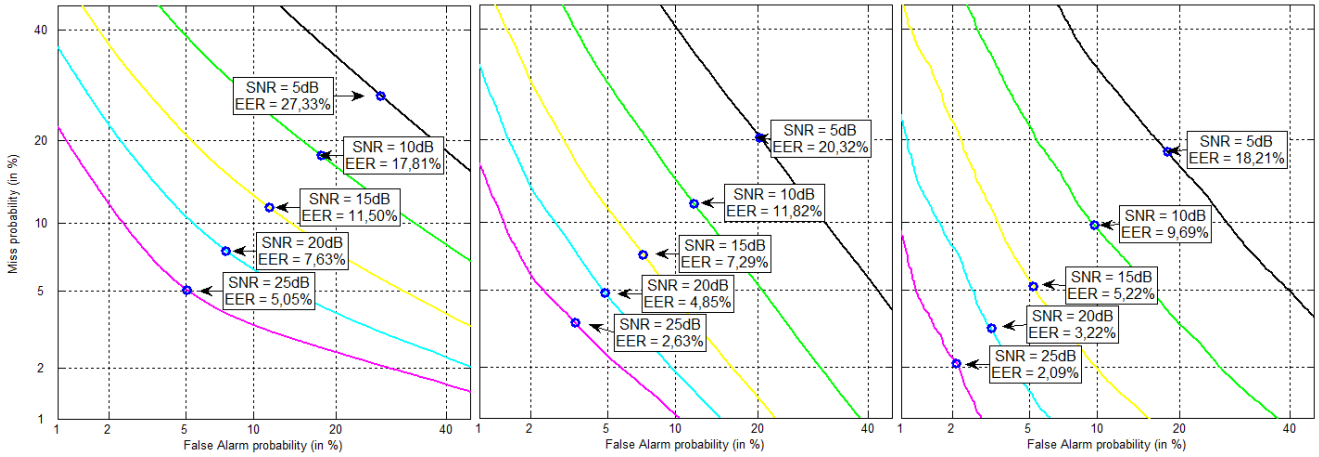

Figure 8. Detection Error Trade-off curves at different SNR without temporal integration (left), with the proposed 1-second median filter on decision output (middle), and with application of our adaptive multi-level segmentation process (right).

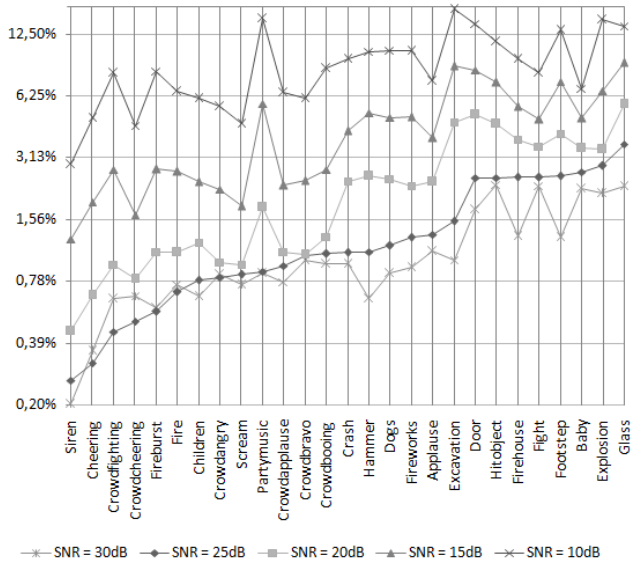

Figure 9. EER performances by event type. For an easier reading, a logarithmic scale is used on ordonates.

for online adaptive integration. 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 automatic tuning of hyperparameters based on automatic analysis of the training signal. A comparative evaluation of different adaptive integration processes should be performed. We will also set up this scheme into a more complete framework including a classification step.

References

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

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

[3] M. Cettolo, M. Vescovi, and R. Rizzi. Evaluation of bic-based algorithms for audio segmentation. *Computer Speech & Language*, 19(2):147–170, 2005.

[4] H. Fletcher and W. Munson. Loudness, its definition, measurement and calculation. *Journal of Acoustical Society of America*, 5:82–108, 1933.

[5] J. Glass and V. Zue. Multi-level acoustic segmentation of continuous speech. In *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, volume 1, pages 429–432, 1988.

[6] J.-L. Husson and Y. Laprie. A new search algorithm in segmentation lattices of speech signals. In *Proceedings of the 4th International Conf. on Spoken Language Processing*.

[7] Internation Telecommunication Union. *Measurement of Audio-Frequency Noise Voltage Level in Sound Broadcasting*. Recommandation, Broadcasting Service, 1986.

[8] International Electrotechnical Commission. *IEC-61672-2 Sound Level Meters - Part 2: Pattern Evaluation Tests*. 2003.

[9] International Organization for Standardization. *ISO-226 Acoustics - Normal Equal-Loudness-Level Contours*. ISO Standards, 2003.

[10] D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J.-F. Serignat. Information extraction from sound for medical telemonitoring. *Information Technology in Biomedicine, IEEE Transactions on*, 10(2):264–274, 2006.

[11] National Institute of Standards and Technology. Det-curve plotting software, information technology laboratory, detware v.2.1.

[12] A. Rabaoui, M. Davy, S. Rossignol, Z. Lachiri, and N. Ellouze. Improved one-class svm classifier for sounds classification. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 117 –122, 2007.

[13] T. Räty. Survey on contemporary remote surveillance systems for public safety. *Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 40(5):493–515, 2010.

[14] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13:1443–1471, 2001.

[15] Sound Ideas. *The Series 6000 "The General" Sound Effect Library*.

[16] M. Tohmé and R. Lengellé. Maximum margin one class support vector machines for multiclass problems. *Submitted to Pattern Recognition Letters*, 2011.

[17] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection and localization for audio-surveillance systems. In *Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, pages 21–26. IEEE Computer Society, 2007.

[18] M. Valera and S. Velastin. Intelligent distributed surveillance systems: a review. *Vision, Image and Signal Processing, IEE Proceedings* -, 152(2):192–204, 2005.