2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance

An ensemble of rejecting classifiers for anomaly detection of audio events

Donatello Conte, Pasquale Foggia, Gennaro Percannella, Alessia Saggese, Mario Vento
University of Salerno
Department of Electronic and Computer Engineering
Via Ponte don Melillo, 1 - 84084 Fisciano (SA), Italy
{dconte, pfoggia, pergen, asaggese, mvento}@unisa.it

Abstract

Audio analytic systems are receiving an increasing interest in the scientific community, not only as stand alone systems for the automatic detection of abnormal events by the interpretation of the audio track, but also in conjunction with video analytics tools for enforcing the evidence of anomaly detection. In this paper we present an automatic recognizer of a set of abnormal audio events that works by extracting suitable features from the signals obtained by microphones installed into a surveilled area, and by classifying them using two classifiers that operate at different time resolutions. An original aspect of the proposed system is the estimation of the reliability of each response of the individual classifiers. In this way, each classifier is able to reject the samples having an overall reliability below a threshold. This approach allows our system to combine only reliable decisions, so increasing the overall performance of the method. The system has been tested on a large dataset of samples acquired from real world scenarios; the audio classes of interests are represented by gunshot, scream and glass breaking in addition to the background sounds. The preliminary results obtained encourage further research in this direction.

1. Introduction

In the recent years audio analytics has emerged more and more as a relevant tool for improving security of the persons and of public and private assets. In fact, in many cases the analysis of the audio signal acquired by one or more microphones deployed into a surveilled area allows to detect abnormal situations that may represent a risk for the public security, more reliably than the video analytics counterpart. The analysis of the audio signal can be used for raising the attention of the surveillance operator on a specific camera, or in conjunction with a video analytics tool in a sort of multiclassifier system, or, still more importantly, for the surveillance of areas where video is not allowed.

For these reasons, in the recent years we have assisted to a growing interest toward these issues. In [1] the authors describe a shot detection system that uses audio information. In a first stage the system segments the audio stream into successive frames of 20 ms and attributes them to the *shot* class or to the *normal* class. The classification is done by a GMM on the basis of features typically used in the context of audio classification as short time energy, MFCC, spectral statistical moments. Then event detection is carried out on a 0.5 second decision window using a Maximum A Posteriori decision rule. Similar approaches based on a GMM classifier have been adopted also in [15, 13, 16]. In particular, Vacher et al. [15] propose a system for scream and glass break detection in indoor applications (apartments) using wavelet based cepstral coefficients. Rouas et al. [13] use energy and MFCC features and test both GMM and SVM for shout detection in outdoor scenarios (railway). They also propose to adopt adaptive thresholding for determining sound activity, with the aim of limiting the false detections. Valenzise et al. [16] discuss a system that discriminates among ambient noise, scream or gunshot, by using two parallel GMM classifiers for identifying screams from noise and gunshots from noise. More recently, Ntalampiras et al. in [10] presented a hierarchical system that classifies the sound as vocalic (normal or screamed speech) or non-vocalic (background environment, gunshot or explosion) event. Based on this decision a different path is chosen to further characterize the audio signal. Depending on the chosen path, ad hoc descriptors are calculated and specific GMM classification stages are activated in order to provide the final audio classification.

The methods in [8, 11] face the problem of abnormal audio event detection adopting a novelty detection approach, with the aim of better dealing with data which differ considerably from those seen by the system during the training phase. In particular, Lecomte et al. in [8] use a One-Class Support Vector Machine (OC-SVM) to model the distribution of the normal sounds and then construct some sets

DOI 10.1109/AVSS.2012.9

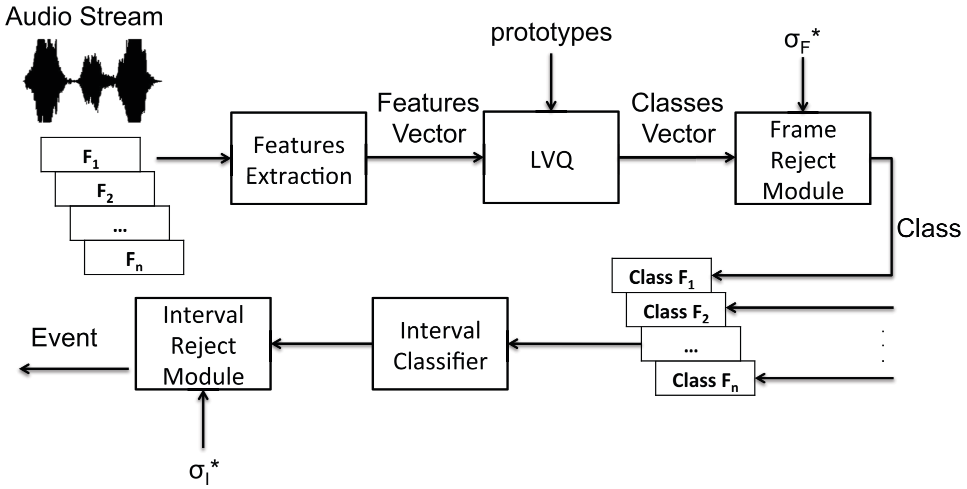

Figure 1. The system architecture of the proposed audio event classification system.

of decision functions that allow to control the trade-off between false-alarm and miss probabilities without re-training the OC-SVM. Similarly, in [11] the authors model normality in the sound track testing GMM, SVM and HMM. In both [8, 11], tests were performed on a large datasets of samples coming from the use of their systems in real contexts for hours or days.

All the considered approaches show very interesting results in the different domains of application demonstrating that the technology is quite mature for the realization of systems to be employed in real environments. Consistently with other approaches in the literature, the method proposed in this paper analyzes the input signal at two resolution levels: at the first level it classifies short segments of few dozens of milliseconds into a set of predefined classes of interest for the application at hand. At the second level, the classification outputs are aggregated into longer time segments, whose duration is comparable to that of the events to detect, typically of the order of seconds. However, the original aspect of the proposed system, that is also the main contribution provided in this paper, is the estimation of the reliability of each output provided by the classifiers of the system: in this way, each classifier is able to reject the samples having an overall reliability below a threshold; this approach allows our detection system to combine only reliable decisions, ignoring outliers that could to false positive detection.

The paper is organized as follows: in Section 2 we describe the architecture of the proposed system, providing also some details about the adopted framework for sample rejection, while in Section 3 we present the dataset used for experimetal validation and the results of the tests. Finally, we draw conclusions and discuss future directions of our research.

2. System Architecture

The proposed system analyze the audio signal at two levels of granularity. At the first level it classifies the time frames, where a time frame is a short segment of the signal whose time span is few dozens of milliseconds. The time frames are partially overlapped. At the second level the proposed system combines the classification outputs obtained on adjacent frames in order to detect events of interest. The analysis at two levels is quite common in the design of audio event classification systems and it is motivated by the simple observation that most of the audio signal descriptors available in the literature are defined on the frame level, while the events of interest are typically characterized by a larger duration.

The system architecture of the proposed method for audio classification is shown in Figure 1. The audio signal is fed to the features extraction module: this module implements the set of features outlined in Table 1. The features, calculated on a frame basis, are used by an LVQ neural network for assigning the audio class (i.e. background noise, scream, gun shot, ...) to the time frame. The frame reject module attributes the sample to the class guessed by the LVQ or rejects it when the reliability is below a threshold σ_F^*. The value of threshold is determined according to the method described in subsection 2.1. The frame reject module requires that the LVQ classifier provided also the reliability together with the guessed class. To this aim we adopted the method defined in [2] for the estimation of the reliability of the frame level classifier[1] ψ^F, which is calculated as a combination of two contributions that accounts for the following situations that can be the cause of unre-

[1] In the following we will denote with superscript I and F those quantities referred to the frame and the interval classifiers, respectively.

liable classifications: (a) the considered sample is significantly different from those present in the dataset, so that the point is located in a region of the feature space far from those occupied by the samples of the training set and associated to the various classes; (b) the point which represents the considered sample in the features space lies where the regions pertaining to two or more classes overlaps. Specifically, we calculated:

$$\psi^F = \min\{\psi^F_a, \psi^F_b\} \tag{1}$$

where ψ^F_a and ψ^F_b accounts for the first and the second causes of unreliability, respectively. The definitions of ψ^F_a and ψ^F_b in case of an LVQ classifier are detailed in [2] and here briefly recalled. The reliability parameter ψ^F_a is defined as:

$$\psi^F_a = \begin{cases} 1 - \frac{O^F_{win}}{O^F_{max}}, & \text{if} \quad O^F_{win} \leq O^F_{max} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where O^F_{win} is the distance of the sample from the closest prototype (associated to the guessed class), while O^F_{max} is the highest value of O^F_{win} among those relative to all the samples of the training set. The reliability parameter ψ^F_b takes into account also the distance O^F_{2win} from the second closest prototype, but associated to a different class with respect to the guessed one. Specifically,

$$\psi^F_b = 1 - \frac{O^F_{win}}{O^F_{2win}} \tag{3}$$

Classification outputs of n consecutive overlapped frames are collected in order to provide the final detection of the events of interest. The value of n is chosen so as to have a time span (hereinafter *interval*) that is significant for the type of events to discriminate (it typically ranges between 250ms and 1s).

The interval classification module computes the reliability using a different definition of O^I_{win}. In fact, it defines the plausibility O^I_i that the sample at hand can be attributed to the i-th class as the ratio between the number of frames in the interval attributed to that class and n. Consequently, in this case $O^I_{win} = \max_i\{O^I_i\}$, so that ψ^I_a and ψ^I_b can be defined as:

$$\psi^I_a = O^I_{win} \tag{4}$$

$$\psi^I_b = O^I_{win} - O^I_{2win} \tag{5}$$

and the overall classification reliability if the interval classifier ψ^I can be measured as

$$\psi^I_b = \min\{\psi^I_a, \psi^I_b\} = O^I_{win} - O^I_{2win} = \psi^I_b \tag{6}$$

Again plausibility information is used by the interval reject module in order to provide the final classification of the interval, so that the label assigned to each interval can be: no event (background noise), event of interest (depending on the applicative scenario it can be scream, glass break, gun shot, explosion, ...), uncertain (rejected sample). The first two labels are assigned to the samples classified with high reliability, while all the unreliable samples fall in the third category.

2.1. Optimal values of the reject threshold

The rationale of the method used in this paper for fixing the reject threshold has been presented in [3]. For the sake of completeness, in the following we will briefly review it.

To this regard, it is assumed that an effectiveness function P, taking into account the requirements of the particular application, evaluates the quality of the classification in terms of correct recognition, misclassification and rejection rates. Under this assumption the optimal reject threshold value, determining the best trade-off between reject rate and misclassification rate, is the one for which the function P reaches its absolute maximum. The requirements of a given application domain are specified by attributing costs to misclassifications, rejects and correct classifications. The cost of an error can be a function of the guess and of the actual class [14]. To operatively define the function P, let us refer to a general classification problem. Suppose that the patterns to be classified can be assigned to one of N+1 classes, labeled with $0, 1, ..., N$. Labels $1, ..., N$ denote the actual classes, while 0 is a fictitious class collecting the rejected patterns. For each actual class i, let us call R_{ii} the percentage of patterns correctly classified, R_{ij} the percentage of patterns erroneously assigned to the class j (with $j \neq i$) and R_{i0} the percentage of rejected patterns.

For the same class i, let R^0_{ii} and R^0_{ij} respectively indicate the percentages of patterns correctly classified and of patterns erroneously assigned to the class j, when the classifier is used at 0-reject. If we assume for P a linear dependence on R_{ii}, R_{ij} and R_{i0}, its expression is given by:

$$P = \sum_{i=1}^{N} C_{ii}(R_{ii} - R^0_{ii}) + \\ - \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} C_{ij}(R_{ij} - R^0_{ij}) - \sum_{i=1}^{N} C_{i0}R_{i0} \tag{7}$$

In other words, P measures the actual effectiveness improvement when the reject option is introduced, with respect to the performance of the classifier at 0-reject. The term C_{ij} denotes the cost of assigning to the class j a pattern belonging to the class i. Note that, if $j = 0$, this is the cost of rejecting a pattern coming from the class i, while, if $j = i$, C_{ij} actually represents the gain associated to a correct classification. Obviously, in order that a rejection be convenient, for each class i, the following relation must

Table 1. The set of features used for describing the audio signal.

Category	Feature name	Reference
Spectral features	Spectral centroid, spectral skewness, spectral kurtosis, spectral slope, spectral decrease, spectrum rolloff	[12][6]
Global Temporal Features	Temporal Decrease, Temporal Centroid,	[12][6]
Energy Feature	TotalEnergy, ERSB	[6] [9]
Instantaneous Temporal Features	Zero-crossing rate	[12] [6]
Perceptual features	Sharpness, Spread, TotalLoudness, Specific Loudness, Normalized Specific Loudness	[12] [7]
Other features	Spectral Crest, Volume	[12]

hold:

$$C_{ij} \geq C_{i0} \quad \forall j \neq 0, j \neq i \tag{8}$$

Since R_{ii}, R_{ij} and R_{i0} depend on the value of the reject threshold σ, P is also a function of σ. Starting from the results presented in [3], it is possible to show that the following relation holds:

$$P(\sigma) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (C_{ij} - C_{i0}) \int_0^{\sigma} D_{ij}(\psi) d(\psi) + \\ - \sum_{i=1}^{N} (C_{ii} + C_{i0}) \int_0^{\sigma} D_{ii}(\psi) d(\psi) \tag{9}$$

where $D_{ii}(\psi)$ and $D_{ij}(\psi)$ (with $j \neq i$) are, respectively, the occurrence density curves of correctly classified and misclassified patterns for the class i as a function of the value of σ. In other words, $D_{ij}(\psi)d(\psi)$ is the fraction of patterns of the class i assigned to class j with a reliability in the interval $[\psi, \psi + d\psi]$.

The optimal value σ^* of the reject threshold σ is the one for which the function P gets its maximum value. In practice, the functions $D_{ij}(\psi)$ are not available in their analytical form and therefore, for evaluating σ^*, they should be experimentally determined in tabular form on a set of labeled patterns, adequately representative of the target domain. The value of σ^* can be then determined by means of an exhaustive search among the tabulated values of $P(\sigma)$.

3. Experimental results

The performance of the proposed system are evaluated on a large dataset of audio samples collected from the Internet or recorded by the authors. The dataset is composed of signals belonging to the following classes: background noise, scream, gun shot, broken glass. The background noise was acquired in indoor and outdoor environments to account for different applicative scenarios. The composition of the dataset is summarized in Table 2. The audio tracks in the dataset are sampled at 8 kHz and mono.

Table 2. Composition of the dataset used for the experimentations.

Class	Time duration
Background noise (BN)	3707,5 secs
Broken glass (BG)	1471,8 secs
Gun shot (GS)	1042,1 secs
Scream (S)	674,3 secs

We partitioned the test database into three equally sized sets preserving the original class distribution: one set (*training set*) was used for training the LVQ neural network, a second set (*validation set*) was used for determining the optimal values of the reject thresholds at the frame and the interval levels, while on the third set (*test set*) we determined the overall performance of the system.

In order to derive the set of the most discriminative features for the problem at hand we performed a preliminary features selection by means of the Weka tool [5]. In particular, we used the *CFS Subset Eval* method [4], which evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them; subsets of features that are highly correlated with the class while having low inter-correlation are preferred. As for the search method, we used the *Best First* strategy. At the end, we individuated the descriptors that are reported in table 3.

The first experiment was aimed at evaluating the impact of the adoption of the reject option after the LVQ classifier at the frame level. The ideal behavior that we would like to obtain after the adoption of this module is that the all samples that are erroneously classified by the LVQ should be rejected, while all the samples that the classifier recognizes correctly should be mantained. In such ideal case the classifier at the interval level aggregates only the samples that at the frame level were attributed to the correct class. Unfortunately, in real cases due to the fact that the distributions of the correctly classified samples and of the misclassified

Table 3. The set of the most discriminants descriptors selected in each category of features.

Category	**Feature name**
Spectral Features	Spectral centroid, spectrum rolloff, spectral pitch, spectral flux
Global Temporal Features	Temporal Decrease, Temporal Centroid
Energy Features	TotalEnergy, ERSB
Instantaneous Temporal Features	Zero-crossing rate
Perceptual features	Spread
Other features	Volume

ones as a function of the classification reliability are usually partially overlapped, it happens that on one side some samples that were correctly classified by the original classifier are rejected and on the other side some samples that were misclassified by the original classifier are not rejected, so they are passed to the next stage as reliable ones.

This behavior can be observed by considering the results reported in Table 4 where we consider the confusion matrix in case of the base LVQ classifier and in Table 5 that accounts for the performance of the classifier with the reject option. The results in the latter table were obtained by setting the cost of the rejections $C_{i0} = 2$, the cost of misclassifications $C_{ij} = 5$, with $i \neq j$, and the gain of correct classifications $C_{ii} = 1$. For the sake of simplicity in this paper we do not consider a dependence of the costs and of the gains by the specific class, even if in real cases such features provided by the adopted rejection framework may be very useful. The choice of the costs used for experimentations is motivated by the fact that at the frame level we are mainly interested to configurations with high values of the cost of the classification errors, while we may accept also relatively high rates of rejections of correctly classified samples. In fact, at this level the primary goal is to obtain the highest percentage of correct classifications even if on a subset of the available samples. The comparison of the results reported in the Tables 4 and 5 shows that the adoption of the reject module after the LVQ classifier allows to reduce in all cases the number of misclassification. This is more evident in the case of the **BG** class that is largely confused with the **BN** class.

In the second experiment we considered the recognition performance of the system at the interval level. The obtained results are reported in Tables 6, 7 and 8. In particular, in Table 6 it is reported the confusion matrix at the interval level when the reject option is not enables at both levels of the system. In this case it is possible to note that the system is able to recognize correctly all the classes with the exception of the **BG** that is confused about one time over four with the **BN** class. We observe a significant improvement of the recognition rates going from the frame level to the interval one that in absolute value is around 10% for all classes. However, if we focus on the results reported in Table 7 obtained when the reject otion is enabled only at the frame level, we can notice a further 10% increase of the accuracy over the **BG** class without any negative impact on the recognition rates of the other classes. Finally, in Table 8 we consider the results obtained by the system at the interval level when the reject option is enabled at both levels. In this case we can notice again that the adoption of this module does not impact over the recognition of the **BN**, **GS** and **S** classes, but only a small fraction of erroneously classified samples from the **BG** class were rejected, suggesting that the use of the reject option is more beneficial if adopted only at the frame level.

Table 4. Confusion matrix of the base LVQ classifier adopted at the frame level on the considered dataset.

	BN	**BG**	**GS**	**S**
BN	91.61%	6.64%	0.50%	1.24%
BG	35.32%	54.93%	4.96%	4.80%
GS	1.66%	4.19%	90.49%	3.66%
S	1.30%	4.35%	4.15%	90.20%

Table 5. Confusion matrix of the LVQ classifier with the reject option adopted at the frame level on the considered dataset. The cells in the rightmost column report the percentage of the rejected samples.

	BN	**BG**	**GS**	**S**	***Rej.***
BN	90.61%	5.87%	0.44%	1.01%	*2.07%*
BG	33.53%	52.08%	4.37%	4.04%	*5.98%*
GS	1.59%	3.43%	89.35%	3.09%	*2.55%*
S	1.12%	3.75%	3.79%	89.26%	*2.08%*

4. Conclusions

In this paper we have proposed a system for automatic recognition of the events of interest from the audio signals acquired by microphones. The main feature of the system is the capability of estimating the reliability of each classification act that is used for rejecting unreliably classified samples. In this way the system takes its decisions by relying only on samples on which it is more confident. This approach allows to improve recognition performance with respect to the base classifier. The system was tested on a dataset of audio events acquired in real operational scenarios showing interesting performance.

Future work will be devoted to a more extensive tests that will consider a larger set of classes of interest (for instance the detection of the aerosol spray for the realization of an audio based detector of graffiti activities), and to the adoption of more sophisticated rules of aggregation of the frame samples at the interval level with the aim of considering variable length intervals that better fit to the typical duration of the events of interest.

Table 6. Confusion matrix at the interval level on the considered dataset when the reject option is not enabled at both the frame and interval level.

	BN	**BG**	**GS**	**S**
BN	100.00%	0.00%	0.00%	0.00%
BG	26.61%	73.39%	0.00%	0.00%
GS	0.00%	0.00%	100.00%	0.00%
S	0.00%	0.00%	0.00%	100.00%

Table 7. Confusion matrix at the interval level on the considered dataset when the reject option is enabled only at the frame level.

	BN	**BG**	**GS**	**S**
BN	100.00%	0.00%	0.00%	0,00%
BG	16.97%	83.03%	0.00%	0.00%
GS	0.00%	0,00%	100.00%	0.00%
S	0.00%	0.00%	0.00%	100.00%

Table 8. Confusion matrix at the interval level on the considered dataset when the reject option is enabled at both the frame and interval level.

	BN	**BG**	**GS**	**S**	***Rej.***
BN	100.00%	0.00%	0.00%	0.00%	*0.00%*
BG	16.61%	81.92%	0.00%	0.00%	*1.48%*
GS	0.00%	0.00%	100.00%	0.00%	*0.00%*
S	0.00%	0.00%	0.00%	100.00%	*0.00%*

References

[1] C. Clavel, T. Ehrette, and G. Richard. Events detection for an audio-based surveillance system. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1306 –1309, july 2005.

[2] C. De Stefano, C. Sansone, and M. Vento. To reject or not to reject: that is the question-an answer in case of neural classifiers. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 30(1):84 –94, feb 2000.

[3] P. Foggia, C. Sansone, F. Tortorella, and M. Vento. Multiclassification: reject criteria for the bayesian combiner. *Pattern Recognition*, 32(8):1435 – 1447, 1999.

[4] M. Hall. Correlation-based Feature Selection for Machine Learning, 1998.

[5] G. Holmes, A. Donkin, and I. H. Witten. WEKA: a machine learning workbench. In *Intelligent Information Systems,1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 357–361, Aug. 2002.

[6] IEEE and ISO/IEC. Multimedia Content Description Interface - Part 4: Audio. *ISO/IEC 42010 IEEE Std 1471-2000 First edition 2007-07-15*, 2001.

[7] M. B. C. J., G. B. R., and B. T. A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–240, 1997.

[8] S. Lecomte, R. Lengelle, C. Richard, F. Capman, and B. Ravera. Abnormal events detection using unsupervised one-class svm - application to audio surveillance and evaluation -. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pages 124 –129, 30 2011-sept. 2 2011.

[9] Z. Liu, Y. Wang, and T. Chen. Audio Feature Extraction and Analysis for Scene Segmentation and Classification. *The Journal of VLSI Signal Processing*, 20(1):61–79, Oct. 1998.

[10] S. Ntalampiras, I. Potamitis, and N. Fakotakis. An adaptive framework for acoustic monitoring of potential hazards. *EURASIP J. Audio Speech Music Process.*, 2009:13:1–13:15, Jan. 2009.

[11] S. Ntalampiras, I. Potamitis, and N. Fakotakis. Probabilistic novelty detection for acoustic surveillance under real-world conditions. *Multimedia, IEEE Transactions on*, 13(4):713 –719, aug. 2011.

[12] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Tech. rep., IRCAM, 2004.

[13] J.-L. Rouas, J. Louradour, and S. Ambellouis. Audio events detection in public transport vehicle. In *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, pages 733 –738, sept. 2006.

[14] C. Sansone, M. Vento, and F. Tortorella. A classification reliability driven reject rule for multi-expert systems. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(6):885–904, 2001.

[15] M. Vacher, D. Istrate, L. Besacier, J. F. Serignat, and E. Castelli. Sound Detection and Classification for Medical Telesurvey. In C. ACTA Press, editor, *Proc. 2nd Conference on Biomedical Engineering*, pages 395–398, Innsbruck, Austria, Feb. 2004.

[16] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti. Scream and gunshot detection in noisy environments. In *Proc. EURASIP European Signal Processing Conference*, Poznan, Poland, 2007.