SCREAM AND GUNSHOT DETECTION IN NOISY ENVIRONMENTS

L. Gerosa, G. Valenzise, M. Tagliasacchi, F. Antonacci, A. Sarti

Dipartimento di Elettronica e Informazione, Politecnico di Milano
Piazza Leonardo da Vinci 32, 20133, Milano, Italy
email: luigi@gerosa.biz, {valenzis, tagliasa, antonacc, sarti}@elet.polimi.it

ABSTRACT

This paper describes an audio event detection system which automatically classifies an audio event as ambient noise, scream or gunshot. The classification system uses two parallel GMM classifiers for discriminating screams from noise and gunshots from noise. Each classifier is trained using different features, appropriately chosen from a set of 47 audio features, which are selected according to a 2-step process. First, feature subsets of increasing size are assembled using filter selection heuristics. Then, a classifier is trained and tested with each feature subset. The obtained classification performance is used to determine the optimal feature vector dimension. This allows a noticeable speed-up w.r.t. wrapper feature selection methods. In order to validate the proposed detection algorithm, we carried out extensive experiments on a rich set of gunshots and screams mixed with ambient noise at different SNRs. Our results demonstrate that the system is able to guarantee a precision of 90% at a false rejection rate of 8%

1. INTRODUCTION

Video-surveillance applications are becoming increasingly important both in private and public environments. As the number of sensors grows, the possibility of manually detecting an event is getting impracticable and very expensive. For this reason, research on automatic surveillance systems has recently received particular attention. In particular, the use of audio sensors in surveillance and monitoring applications has proved to be particularly useful for the detection of events like screams or gunshots [2][11]. Such detection systems can be efficiently used to signal to an automated system that an event has occurred with high probability and, at the same time, to enable further processing like automatic video-camera steering.

Audio-based surveillance stems from the field of automatic audio classification and matching. Traditional tasks in this area are speech/music segmentation and classification [7][9] and audio retrieval [14]. More recently, specific works covering the detection of particular classes of events for multimedia-based surveillance have been developed. For example, detection systems specifically designed for impulsive sound recognition consist of a segmentation step, in which the presence of an event is detected, followed by a classification step, which refines the result assigning a class label to the event. The results reported in [4] show that these systems fail under real-world conditions reaching less than 50% accuracy at 0 dB SNR. In the SOLAR system presented in [6], the segmentation step is avoided by decomposing audio tracks into short, overlapping audio windows. For each window, a set of 138 features is extracted and evaluated by a series of boosted decision trees. Though efficient in real time computations, the SOLAR system suffers from large differences in classification accuracy from class to class.

More recent works have shown that a hierarchical classification scheme, composed by different levels of binary classifiers, generally achieves higher performance than a single-level multi-class classifier. In [1] a hierarchical set of cascaded Gaussian Mixture Models (GMM) is used to classify 5 different sound classes. Each GMM is tuned using only one feature from a feature set including both scalar features (e.g. ZCR - Zero Crossing Rate) or vector features (e.g. Linear Log Frequency Cepstral Coefficients). Reported results show that the hierarchical approach yields accuracies from 70 to 80% for each class, while single level approaches reach high accuracies for one class but poor results for the others.

The hierarchical approach has also been employed in [11] to design a specific system able to detect screams/shouts in public transport environments. After a preliminary segmentation step, a set of perceptual features such as MFCC (Mel-Frequency Cepstral Coefficients) or PLP (Perceptual Linear Prediction) coefficients are extracted from audio segments and used to perform a 3-levels classification. First, the audio segment is classified either as noise or non-noise; second, if it is not noise, the segment is classified either as speech or not speech; finally, if speech, it is classified as a shout or not. The authors have tested this system using both GMMs and Support Vector Machines (SVMs) as classifiers, showing that in general GMMs provide higher precision. A different technique is used in [2] to detect gunshots in public environments. In this work, the performance of a binary gunshot/noise GMM classifier is compared to a classification scheme in which several binary subclassifiers for different types of firearms are run in parallel. A final binary decision (gunshot/noise) is taken evaluating the logical OR of the results of each classifier. In this way, the false rejection rate of the system is reduced by a 50% on average with respect to the original binary classifier.

In this paper we propose a system that is able to accurately detect two types of audio events: screams and gunshots. Our approach is different from the previous works in the following aspects. First, we extract a larger set of features, including some descriptors like spectral slope and periodicity, and innovative features like correlation roll-off and decrease. To the authors' knowledge, these features have never been used for the task of sound-based surveillance. We show that they provide a significant performance gain. Second, we provide an exhaustive analysis of the feature selection process, mixing the classical filter and wrapper feature selection approaches. In the rest of the paper, the audio detection system is described in detail. In Section 2 the families of features used in the system are presented. Section 3 describes the feature selection process, Section 4 details the classification architecture, which is composed by two parallel GMM classifiers, while Section 5 provides the results of classification using the selected features. In Section 6 we discuss some final considerations and future works.

2. AUDIO FEATURES

A considerable number of audio features have been used for the tasks of audio analysis and content-based audio retrieval. Traditionally, these features have been classified in *temporal features*, e.g. Zero Crossing Rate (ZCR); *energy features*, e.g. Short Time Energy (STE); *spectral features*, e.g. spectral moments, spectral flatness; *perceptual features*, e.g. loudness, sharpness or Mel Frequency Cepstral Coefficients (MFCCs). In this work, we have chosen to discard audio features which are too sensitive to the SNR conditions, like STE and loudness. In addition to the traditional features listed above, we employ some other features which have

The work presented was developed within VISNET II, a network of excellence funded by the European Commission (http://www.visnet-noe.org)

not been used before in similar works, such as *spectral distribution* (spectral slope, spectral decrease, spectral roll-off) and *periodicity* descriptors. In this paper we also introduce a few innovative features based on the *auto-correlation* function: correlation roll-off, correlation decrease and correlation slope.

These features are similar to spectral distribution descriptors (spectral roll-off, spectral decrease and spectral slope), but, in lieu of the spectrogram, they are computed starting from the auto-correlation function of each frame. The goal of these features is to describe the energy distribution over different time lags. For impulsive noises, like gunshots, much of the energy is concentrated in the first time lags, while for harmonic sounds, like screams, the energy is spread over a wider range of time lags. Features based on the auto-correlation function are labeled in two different ways, filtered or not filtered, depending on whether the autocorrelation function is computed, respectively, on a band-pass filtered version of the signal or on the original signal. The rationale behind the filtering approach is that much of the energy of some signals (e.g. screams) is distributed in a relatively narrow range of frequencies; thus the autocorrelation function of the filtered signal is much more robust to noise. In this paper, the limits of the frequency range for filtering the autocorrelation function have been fixed to $300-800$ Hz: experimental results have shown that most of the energy of the screams events is concentrated in this frequency range.

To evaluate the discrimination power of the selected features, two feature sets have been created. The first set contains 36 traditional features (ZCR, Spectral Flatness, spectral moments, 30 MFCC coefficients). The second set is composed by the the previous features plus the following descriptors: periodicity, spectral distribution descriptors and correlation distribution descriptors, for a total size of 47 features. Table 1 lists the feature set composition. All the features are extracted from 23ms analysis frames (at a sampling frequency of 22050 Hz) with $1/3$ overlap.

3. FEATURE SELECTION

Starting from the full set of 47 features, we can build a feature vector of any dimension l, $1 \leq l \leq 47$. It is desirable to keep l small in order to reduce the computational complexity of the feature extraction process and to limit the over-fitting produced by the increasing number of parameters associated to features in the classification model.

In previous works on audio surveillance systems, the "best" feature vectors have been empirically determined [6], without adopting an objective performance metrics. In this paper, we discuss an automatic procedure aimed at selecting the best feature set, according to some objective performance indicator. To this end, two main feature selection approaches have been discussed in literature [3]. In the *filter* method, the feature selection algorithm is independent of any classifier, filtering out features that have little chance to be useful in the analysis of data. The filter methods are based on performance evaluation metrics calculated directly from the data, without direct feedback from a particular classifier used. The second approach, known as *wrapper* approach, consists of evaluating a feature vector on the basis of classification results, obtained using that specific subset of features. Therefore, these methods exploit some form of feedback provided by the classifier (e.g. accuracy). They tend to outperform filter methods, but at a much higher computational load.

#	Feature Type	Features	Ref.
1	Temporal	ZCR	[7]
2-6	Spectral	4 spectral moments + SFM	[8]
7-36	Perceptual	30 MFCC	[12]
37-39	Spectral distribution	spectral slope, spectral decrease, spectral roll-off	[8]
40-47	Correlation-based	(filtered) periodicity, (filtered) correlation slope, decrease and roll-off	[7][8]

Table 1: Audio features used for classification.

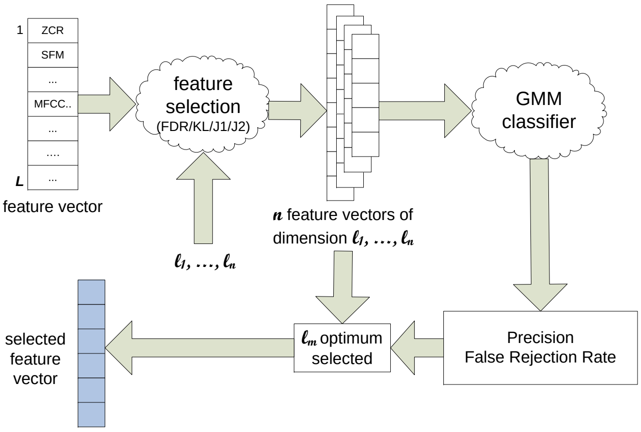


Figure 1: Hybrid feature selection system.

The feature selection process adopted in this work is a hybrid filter/wrapper method. First, a feature subset of size l is assembled from the full set of features according to some class-separability measure and a heuristic search algorithm, as detailed in Section 3.1. The so-obtained feature vector is evaluated by a GMM classifier, which returns some classification performance indicator related to that subset (this procedure is explained in Section 3.2). Repeating this procedure for different l's, one can choose the feature vector dimension that optimizes the desired target performance (see Figure 1). In other words, the hybrid approach splits the problem of feature selection in *two* subproblems: the choice of the feature subset *content*, performed with a filter technique, and the selection of feature vector *dimension*, which is carried out in a wrapper fashion. This combined approach allows a considerable speedup in terms of resources needed for computation w.r.t. a pure wrapper approach, while giving good results for what concerns the overall classification performance.

3.1 Selection of a Feature Vector given a size l

Ideally, the problem of selecting a subset of l features out of the m originally available requires to evaluate all the $\binom{m}{l}$ possible combinations of feature vectors of size l. In practice, the computational complexity is too high. Therefore heuristic methods are used to explore the feature space, searching for a (locally) optimal feature vector. There are two kinds of search algorithms [13]: *scalar* methods, which are based on criteria evaluating the class separability of *individual* features, and *vectorial* methods, which are based on criteria evaluating the class separability of *a vector* of features.

3.1.1 Scalar Selection

In this work, we adopt a feature selection procedure described in [13]. The core idea of this technique consists in building up a feature vector choosing the features that best discriminate the different classes, while at the same time minimizing the correlation between selected features. The method builds a feature vector iteratively, starting from the most discriminating feature and including at each step k the feature $\hat{r}$ that maximizes the following function:

$$J(r) = \alpha_1 C(r) - \frac{\alpha_2}{k-1} \sum_{i \in \mathscr{F}_{k-1}} |\rho_{ri}|, \text{ for } r \neq i. \tag{1}$$

In words, Eq. 1 says that the feature to be included in the feature vector of dimension k has to be chosen from the set of features not yet included in the feature subset $\mathscr{F}_{k-1}$. The objective function is composed of two terms: $C(r)$ is a class separability measure of the rth feature, while ρ_{ij} indicates the cross-correlation coefficient between the ith and jth feature. The weights α_1 and α_2 determine the relative importance that we give to the two terms. In this paper, we use either the Kullback-Leibler divergence (KL) or the Fisher Discriminant Ratio (FDR) to compute the class separability $C(r)$ [13].

3.1.2 *Vectorial Selection*

The vectorial feature selection is carried out using the *floating search* algorithm [10]. This procedure builds a feature vector iteratively and, at each iteration, reconsiders features previously discarded or excludes features selected in previous iterations from the current feature vector. Though not optimal, this algorithm provides better results than scalar selection, but with an increased computational cost. The floating search algorithm requires the definition of a vectorial class separability metrics. In the proposed system, we use either one of the following objective metrics [13]:

$$J_1 = \frac{\mathrm{trace}(S_m)}{\mathrm{trace}(S_w)} \tag{2}$$

$$J_2 = \frac{\det(S_m)}{\det(S_w)} \tag{3}$$

where S_w is the *within-class* scatter matrix, which carries information about *intra*-class variance of the features, while $S_m = S_w + S_b$ is the *mixture* scatter matrix; S_b, the *between-class* scatter matrix, gives information about *inter*-class covariances.

3.2 Selection of the Feature Vector Dimension l

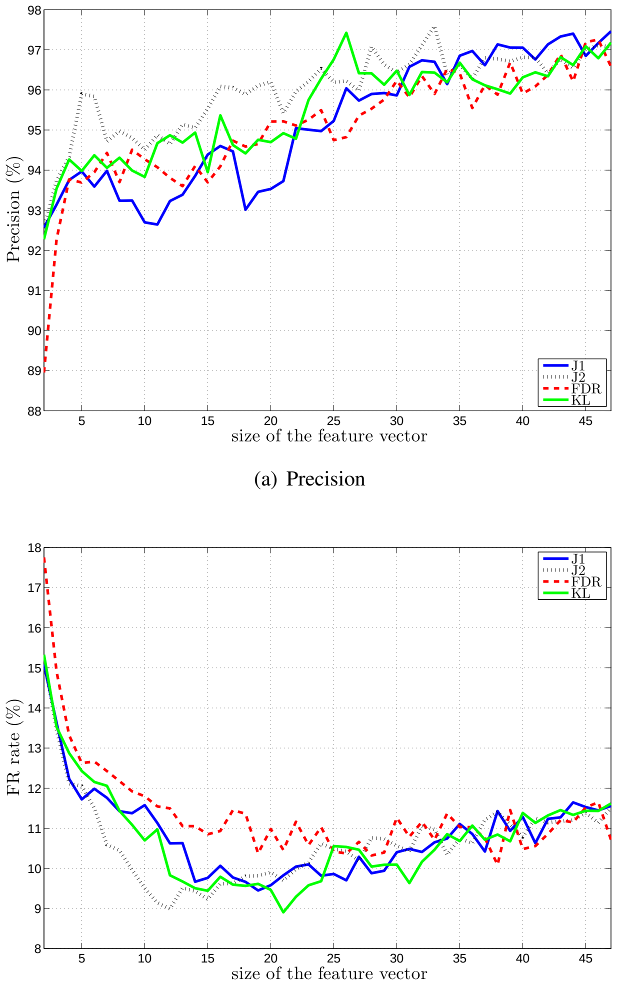
The optimal vector dimension is determined using a wrapper approach. For each dimension l, the aforementioned feature selection algorithm determines the best feature subset for either gunshot/noise or scream/noise classification; the performance of classification using this feature vector are evaluated using the GMM classifier described in the next section. The two performance indicators we take into consideration are the precision and the false rejection rate (FR), defined as follows:

$$\text{precision} = \frac{\text{number of events } \textit{correctly} \text{ detected}}{\text{number of events detected}} \tag{4}$$

$$FR = \frac{\text{number of events } \textit{not} \text{ detected}}{\text{number of events to detect}}, \tag{5}$$

where the term “event” denotes either a scream or a gunshot, depending on which of the two binary classifiers we are considering. The rationale behind the choice of precision and false rejection rate as performance metrics is that in an audio-surveillance system the focus is on minimizing the number of events “missed” by the control system, while at the same time keeping as small as possible the number of false alarms.

We evaluate the precision and false rejection rate for feature vectors of any dimension l. Figure 2 shows how the performance vary as l increases, for the case of scream events (analogous results are obtained with gunshot samples). From these graphs, it is clear that good performance may be obtained with a small number of features, while increasing l above a certain dimension $\hat{l}$ (e.g. 12 in the case of screams as can be argued by figure) performance does not improve significantly. In this work, $\hat{l}$ has been chosen empirically by inspection of the graphs shown in Figure 2. More formal, automatic criteria may be formulated to take in consideration how much of the overall performance is reached with each dimension, weighted by the number of features used. This kind of trade-off optimization will be investigated in a further work.



(a) Precision

(b) False Rejection Rate

Figure 2: Classification precision and false rejection rate of scream with increasing feature vector dimension l.

4. CLASSIFICATION

The event classification system is composed by two Gaussian Mixture Model (GMM) classifiers that run in parallel to discriminate, respectively, between screams and noise, and between gunshots and noise (see Figure 3). Each binary classifier is trained separately with the samples of the respective classes (gunshot and noise, or scream and noise), using the Figueiredo and Jain algorithm [5]. This method is conceived to avoid the limitations of the classical Expectation-Maximization (EM) algorithm for estimating the parameters of a mixture model: through an automatic “component annihilation” procedure, the Figueiredo-Jain algorithm automatically reduces the number of components of the mixture according to an information-theoretic criterion. This way, the issue of selecting the number of components and the problem of determining adequate initial conditions are ruled out; furthermore, singular estimates of the mixture parameters can be automatically avoided by the algorithm.

For the testing step, each frame from the input audio stream is classified *independently* by the two binary classifiers. The decision that an event (scream or gunshot) has occurred is then taken by computing the logical OR of the two classifiers.

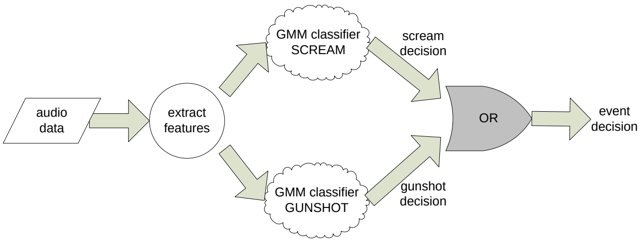


Figure 3: Classification scheme.

5. EXPERIMENTAL RESULTS

In our simulations we have used audio recordings taken from movies soundtracks and internet repositories. Some screams have been recorded live from people asked to shout into a microphone. Finally, noise samples have been recorded live in a public square of Milan.

Audio signals of gunshot and scream classes have been mixed with ambient noise according to some prefixed SNR. We have tested the system with two kinds of experiments. In the first experiment, we want to assess the performance improvements given by the introduction of the additional features described in Section 2. In the second experiment, we test the system performance under different SNR conditions of training and test sequences.

5.1 Performance gain with new features

In this experiment we evaluate the performance gain of the classifier obtained adding the new features described in Section 2, using k-fold cross validation, with $k = 10$. Figure 4 compares precision and FR rate for scream/noise discrimination, using 2, 5, 10, 15 and 20 features, chosen from the set of standard 36 features, denoted as the "old" feature set, and the set of 47 features, called the "new" feature set. In both cases, we have selected features with the vectorial J2 criterion, which turns out to be the best heuristics from Figure 2. The performance metric used is *accuracy*, defined as the number of correct detections over the total number of test samples. Analogous results may be produced for the case of gunshot/noise classification. It must be pointed out that the feature vectors assembled by the vectorial selection algorithm contain mainly some of the additional features presented before (e.g., the feature vector of size 5 in figure contains, among other descriptors, periodicity, spectral roll-off and correlation decrease).

5.2 Effects of SNR on performance

This experiment aims at verifying the effects of the noise level on the training and test sets. We have added noise both to the audio events of the training set and to the audio events of the test set, changing the SNR from 0 to 20dB, with a 5dB step. Making the cross-product of the possible SNR values for training and test sequences, we have built 5^2 classification problems. The performance indicators we have used in this test are the false rejection rate, defined in (5), and the false detection rate (FD), defined as follows:

$$FD = \frac{\text{number of detected events that were actually noise}}{\text{number of noise samples in the test set}}, \quad (6)$$

where, as usual, an event could be both a scream or a gunshot. The results for both scream and gunshot classification are reported in Figure 5. As expected, performance degrades noticeably as the SNR of both training and test sequences decreases. In other words, as the energy of the event is decreased compared to the energy of background noise, the results in terms of false rejection rate become poor. In particular, training database with high SNR (e.g. 20dB) provide good performance in terms of low false detection rate, while if the training samples are highly corrupted by noise (e.g. 0dB SNR), the false detection rate tends to grow up. On the other hand, increasing the SNR of test sequence brings down the curves in terms of false rejection rate; in other words the number of events missed by the classifier grows as the noise level of tested samples increases. A combination of high training SNR and low testing SNR for the case of gunshot/noise discrimination can dramatically deteriorate the false rejection rate, as illustrated in part (a) of Figure 5: with training database at 20dB SNR and test sequences at 0dB SNR, the system is able to identify only about 15% of the actual events occurred. This is due to the noisy nature of gunshots, which at low SNR are easily confused with ambient noise. It must be said, however, that in realistic conditions it is quite improbable that a weapon firing in the range of a video-surveillance system produce a sound with such a low SNR.

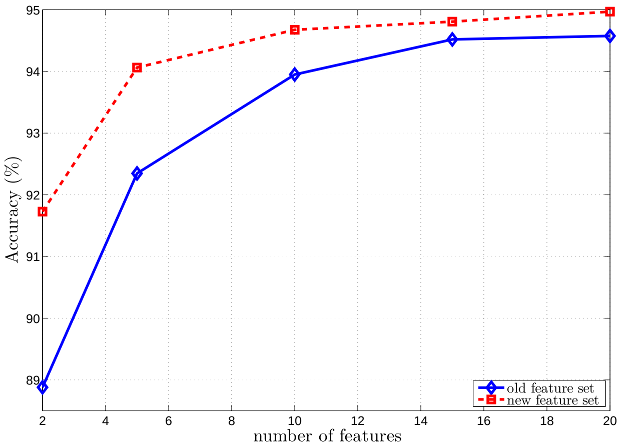


Figure 4: Comparison of accuracy with different feature vector dimensions for scream/noise discrimination. The selection algorithm is vectorial J2.

This experiment illustrates the trade-off existing between false rejection and false detection rate. According to the desired performance of the system, one should choose the appropriate SNR for the training database.

5.3 Combined system

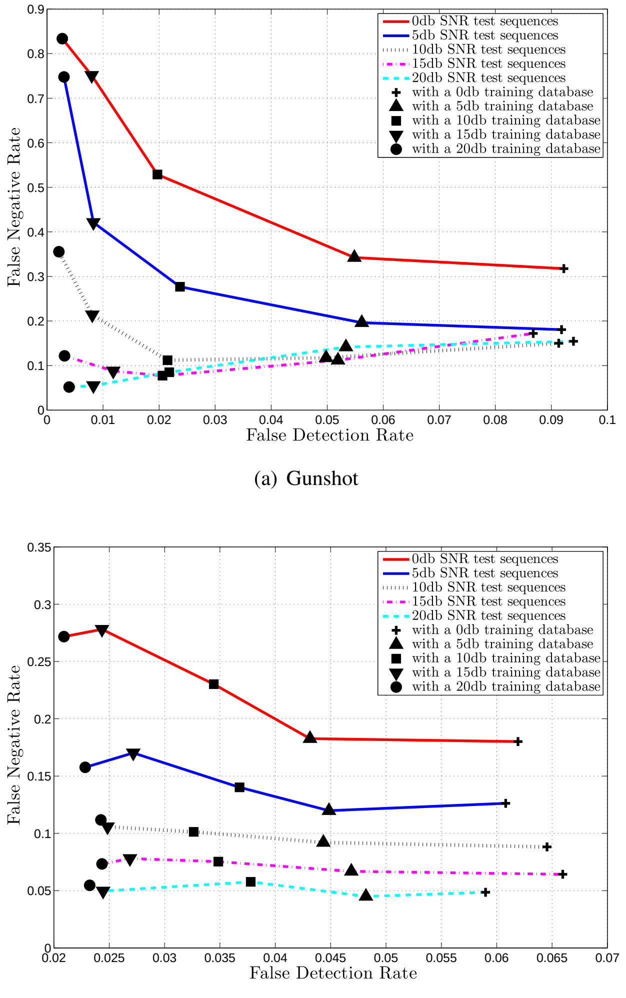
Putting together the scream/noise classifier and the gunshot/noise classifier we can yield a precision of 90% with a false rejection rate of 8%, using scream samples at 10dB SNR and gunshot samples at 15dB SNR. We have used a feature vector of 12 features, selected with the J2 heuristic, for scream/noise classification, and a feature vector of 10 features, selected with the J1 criterion, for gunshot/noise classification. The two feature vectors are reported in Table 2.

6. CONCLUSIONS

In this paper we analyzed a classification system able to detect events such as gunshots and screams in noisy environments. We have considered a set of audio features larger than the sets usually adopted in this kind of tasks. We have provided a method for choosing the composition of the feature vector and for evaluating his dimension mixing filter selection criteria and wrapper validation of results. This approach allows to reduce considerably the dimensionality of the problem, producing, with a small computational cost, a feature vector which gives acceptable results. We have drawn the attention on the necessary trade-off between false rejection and false detection rate, testing the system under very noisy conditions. Without using features directly related to the local energy of signals such as short time energy, we are able to obtain an accuracy of 90% and

#	Gunshot/Noise classifier	Scream/Noise classifier
1	Spectral Centroid	SFM
2	MFCC 1	Spectral Skewness
3	MFCC 2	MFCC 2
4	MFCC 3	MFCC 3
5	MFCC 11	MFCC 9
6	MFCC 28	MFCC 12
7	MFCC 29	(filtered) periodicity
8	MFCC 30	correlation decrease
9	(filtered) periodicity	spectral slope
10	ZCR	correlation slope
11		spectral decrease
12		periodicity

Table 2: Feature vectors used in the combined system



(a) Gunshot

(b) Scream

Figure 5: False rejection rate as a function of false detection rate for various SNR training database and test sequences. $\hat{l}$ is 12 for scream events and 10 for gunshot events.

a false rejection rate of 8% with the combined system.

Future work will be dedicated to the formalization of feature dimension selection, by formulating a trade-off optimization problem which optimizes simultaneously different classification and computational performance metrics.

REFERENCES

[1] P.K. Atrey, N.C. Maddage, and M.S. Kankanhalli. Audio Based Event Detection for Multimedia Surveillance. *Acoustics, Speech, and Signal Processing, 2006. ICASSP-97., 2006 IEEE International Conference on*, 2006.

[2] C. Clavel, T. Ehrette, and G. Richard. Events Detection for an Audio-Based Surveillance System. *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 1306–1309, 2005.

[3] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1(3):131–156, 1997.

[4] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini. Automatic sound detection and recognition for noisy environment. *European Signal Processing Conference (EUSIPCO), Tampere, Finlande*, pages 1033–1036, 2000.

[5] MAF Figueiredo and AK Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[6] D. Hoiem, Y. Ke, and R. Sukthankar. SOLAR: Sound Object Localization and Retrieval in Complex Audio Environments. *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 5, 2005.

[7] L. Lu, H.J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. *Speech and Audio Processing, IEEE Transactions on*, 10(7):504–516, 2002.

[8] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO Project Report*, 2004.

[9] J. Pinquier, J.L. Rouas, and R. Andre-Obrecht. Robust Speech/Music Classification in Audio Documents. *International Conference on Spoken Language Processing, ICSLP*, 5:10–15, 2002.

[10] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

[11] J.L. Rouas, J. Louradour, and S. Ambellouis. Audio Events Detection in Public Transport Vehicle. *Proc. of the 9th International IEEE Conference on Intelligent Transportation Systems*, 2006.

[12] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiøler. Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. In *Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR)*, 2006.

[13] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 2006.

[14] T. Zhang and C.C.J. Kuo. Hierarchical system for content-based audio classification and retrieval. *Conference on Multimedia Storage and Archiving Systems III, SPIE*, 3527:398–409, 1998.