World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:3, No:11, 2009

Using HMM-based Classifier Adapted to Background Noises with Improved Sounds Features for Audio Surveillance Application

Asma Rabaoui, Zied Lachiri, and Noureddine Ellouze

***Abstract*—**Discrimination between different classes of environmental sounds is the goal of our work. The use of a sound recognition system can offer concrete potentialities for surveillance and security applications. The first paper contribution to this research field is represented by a thorough investigation of the applicability of state-of-the-art audio features in the domain of environmental sound recognition. Additionally, a set of novel features obtained by combining the basic parameters is introduced. The quality of the features investigated is evaluated by a HMM-based classifier to which a great interest was done. In fact, we propose to use a *Multi-Style* training system based on HMMs: one recognizer is trained on a database including different levels of background noises and is used as a universal recognizer for every environment. In order to enhance the system robustness by reducing the environmental variability, we explore different adaptation algorithms including Maximum Likelihood Linear Regression (MLLR), Maximum A Posteriori (MAP) and the MAP/MLLR algorithm that combines MAP and MLLR. Experimental evaluation shows that a rather good recognition rate can be reached, even under important noise degradation conditions when the system is fed by the convenient set of features.

***Keywords*—**Sounds recognition, HMM classifier, *Multi-style* training, Environmental Adaptation, Feature combinations.

I. Introduction

Environmental sounds were described as the sounds which fill our everyday acoustic environment [1]. The panoply of environmental sounds is vast, it includes the sounds generated in domestic, business, and outdoor environments. Recently, some efforts have been directed towards systems capable of detecting and classifying environmental sounds [1], [2], [3]. For example, a system able to recognize indoor environmental sounds can offer concrete potentialities for surveillance and security applications. Furthermore, these functionalities can also be used in portable tele-assistive devices, to inform disabled and elderly persons affected in their hearing capabilities about relevant environment sounds (warning signals, etc.).

In the field of environmental sound recognition, many previous works [1], [2], [4], [5] have focused on recognizing single sound events. Few studies have been reported concerning the broader task of automatic sound recognition when many sound classes are considered. Early works in this vein proposed systems devoted to specific tasks such as sound alerting aid [6], considering some classes of stationary signals (bells, phone, door rings, etc.) and using standard statistical classification procedures based on spectral and temporal features. Later, new classifiers were explored using Hidden Markov Models [7] and Neural Networks [8].

Many other works deal with stationary sounds[1]. An extensive thesis was published by Couvreur [1], including a complete bibliographic review of the statistical sound recognition domain[2].

In [4], the author used various pattern recognition frameworks[3] to design noise classification algorithms. In [13], [3] several audio signal analysis methods was evaluated using several classifiers and a great interest was devoted to impulsive sound analysis techniques, including traditional time-frequency transformations, speech-typical coefficients and psycho-acoustical features.

In [14], the author describes a coherent framework for understanding the perceptual organization of sounds. This seminal work has stimulated much interest in computational studies of hearing. Such studies are motivated in part by the need for practical sound separation systems and sound source recognition [15], which have many applications including noise-robust automatic speech recognition [4] and automatic music transcription [16]. This emerging field has become known as computational auditory scene analysis (CASA).

Among previous CASA works, we emphasize the emerging research in [17] in the field of speech and speaker recognition, which demonstrates that non-stationary (time-frequency) techniques can be applied to sound classification and can produce good results. Therefore, in [18], [19], [20] both stationary frequency-based techniques and non-stationary time-frequency-based feature extraction techniques were tested in combination with several common classification techniques[4] for their suitability to environmental sound recognition.

Generally, the developed Automatic Sounds Recognition (ASR) systems were very sensitive to variations between training and testing conditions, whether these variations were related to changes in acoustic environment or incorrect modelling assumptions. Hence, to successfully develop ASR ap-

Asma Rabaoui and Noureddine Ellouze are with the Unité de recherche Signal, Image et Reconnaissance des formes, ENIT, BP 37, Campus Universitaire, 1002 le Belvédère, Tunis Tunisia. emails: Asma.Rabaoui@enit.rnu.tn, N.ellouze@enit.rnu.tn

Zied Lachiri is with the Département Physique et Instrumentation, INSAT, BP 676, Centre Urbain Cedex, 1080, Tunis Tunisia. email: zied.lachiri@enit.rnu.tn

[1]For example, ringing sounds classification [9], helicopter noise identification [10], and recognition of music instruments [11] or music types [12].

[2]It introduces three classifiers for use in separate noise event recognition (car, truck, airplane etc.).

[3]such as the Quadratic Gaussian Classifier (QGC), Least-Square Linear Classifier (LS-LC), Nearest-Neighbor Classifier (NNC), and Decision Tree Classifier (DTC).

[4]traditionally used in speech or in musical instrument recognition.

plications, it is crucial to take into account such discrepancies. This can be achieved using different kinds of techniques [21] aiming essentially at finding robust and invariant signal features, improving the modelling techniques, modifying recognition parameters or features using adaptation or compensation techniques [22], and using robust decision strategies [23].

Various environment-independent (EI) sounds recognition systems have widely been studied in recent years because they show good performance on average owing to their capability of including a wide variety of environment individualities. However, their performance is still lower than that of well trained environment-dependent (ED) sounds recognition systems. Environment adaptation technique has been one breakthrough regarding this problem, and has been applied alongside EI sounds recognition systems.

In this paper, the developed system is intended for the recognition of limited classes of environmental sounds and is motivated by a practical surveillance application. That system which is able to classify a number of different sounds found in environments typical to daily life seems to be better accepted by humans than video camera monitoring. This work forms part of a larger investigation into the integration of sound surveillance in a monitoring application.

Firstly, we treat the applicability of a range of audio features in the domain of environmental sound classification. Therefore, we propose to use a set of dedicated audio features, composed of several classical ones together with original wavelet-based features. The efficiency of this set of audio features as well as the way they are combined is evaluated and compared to established, standard features. The performance of our technique is evaluated on a data set of sounds collected from the commercial databases [24], [25], which include sounds ranging from screams to explosions, gun shots or glass breaks. The quality of the features is examined with the HMM-based classifier. Moreover, the paper focuses on environment adaptation of the acoustic Hidden Markov Models [26] by applying a particular training mode called the *Multi-Style* training. The objective of acoustic model adaptation techniques is to derive a new set of acoustic models from the reference models given some adaptation data reflecting test acoustic conditions.

The remainder of this paper is organized as follows. Section II gives an overview of the HMMs-based sound classifier: pre-processing and baseline classifier. Environmental adaptation of HMMs to real world background noises is presented in Section III. Experimental set-up and results are provided in Section IV. Section V concludes the paper with a summary and discussion.

II. A HMM-BASED SOUND CLASSIFIER

A. Notations

Let $\mathcal{S}$ be the set of sounds, shared in N classes denoted $\mathcal{S}_1, \ldots, \mathcal{S}_N$. Each class contains m_i training sounds and m'_i adaptation sounds, $i = 1, \ldots, N$. Sound $\#j$ in class $\mathcal{S}_i$ is denoted $\mathbf{s}_{i,j}$, ($i = 1, \ldots, N, j = 1, \ldots, m_i + m'_i$). Generally, the pre-processor converts a recorded acoustic signal $\mathbf{s}_{i,j}$ into a time-localized or a frequency-localized representation. Such representations are obtained by splitting the signal $\mathbf{s}_{i,j}$ into $T_{i,j}$ overlapping short frames and computing a vector of features $x_{t,i,j}$, $t = 1, \ldots, T_{i,j}$ with dimension d which characterize each frame.

B. Pre-processing: Features extraction

The features extraction is an important part of a recognizer. If the features are ideally good, the type of classification architecture won't have much importance. On the apposite, if the features can't discriminate between the concerned classes, no classifier will be efficient. Ideally good features should present the following properties:

- they have to emphasize the differences between classes.
- they have to be robust to noise disturbance, preserving the class separability as far as possible.
- a high correlation between features should be avoided as much as possible.

In this paper, we consider environmental sounds, thus features initially designed for speech seem well adapted. However, environmental sounds may differ significantly from speech, thus we additionally consider features that take care of the possible high non-stationarity of the sounds. Thus, the features selected include those derived from the Discrete Fourier Transform (DFT), the Discrete Cosine Transform (DCT) and the Discrete Wavelet Transform (DWT). The advantage of DFT and DCT is that a few coefficients suffice to represent most of the original signal. However, we note that the DWT takes the original signals in time/space domain and transforms them into time/frequency or space/frequency domain, thus keeping the time variable in a natural way.

1) Time-domain features:

- The Zero-Crossing Rate (ZCR) is defined as the number of times the sign of a time series s_k changes within a frame. It roughly indicates the frequency that dominated during that frame. The ZCR is closely related to the spectral centroid (see below) as they both measure a unique frequency over a frame. It is defined for the frame k with the length L as:

$$ZCR(k) = \frac{1}{L}\sum_{\tau=1}^{L} |\text{sign}(s_k(\tau+1)) - \text{sign}(s_k(\tau))| \quad (1)$$

- The short-time average energy (referred to as "Energy" in the following) is the energy of a frame:

$$\text{Energy}(k) = \frac{1}{L}\sum_{\tau=0}^{L-1} |s_k(\tau)|^2 \quad (2)$$

2) Frequency-domain features:

- The Spectral Centroid (SC) represents the balancing point of the spectral power distribution. It is calculated as the average of the frequencies, weighted by the amplitudes. This is the first moment of the spectrum with respect to frequency. SC is commonly associated with the measure of brightness of a sound. The individual centroid of a spectral frame is defined as:

$$SC(k) = \frac{\sum_{i=1}^{N-1} iS_k[i]}{\sum_{i=1}^{N-1} S_k[i]} \quad (3)$$

For the k^{th} frame, $S_k[i]$ is the magnitude corresponding to the i^{th} frequency and N is the length of the Discrete Fourier Transform (DFT).

- The Spectral Roll-off point (SRF) measures the frequency below which a certain amount of the power spectrum lies. This feature is related to the spectral skewness and it changes for sounds with different frequency ranges. It is calculated by summing up the power spectrum samples until the desired percentage or threshold (referred to as TH below) of the total energy is reached. Considering the DFT of a frame, the SRF is defined as:

$$SRF(k) = \max\{I \backslash \sum_{i=0}^{I} |S_k(i)|^2 < TH \sum_{i=0}^{N} |S_k(i)|^2\} \tag{4}$$

where TH is between 0 and 1. The commonly used value is 0.93.

3) Linear Prediction, Perceptual Linear prediction and cepstral features: These features are used to describe the spectral shape of a signal.

- The Mel-Frequency Cepstral Coefficients (MFCCs) are extracted by applying the discrete cosine transform to the log-energy outputs of mel-scaling filter-bank [27].
- Linear Prediction Cepstral Coefficients (LPCCs) are extracted using the autocorrelation method [28]. Given the linear predictive coefficients $a_k, k = 1, \ldots, N$, the Linear Predictive Cepstral Coefficients (LPCCs) are determined by the following recursive relationship:

$$\begin{cases} c_1 = a_1, \\ c_n = \sum_{k=1}^{n-1}(1 - \frac{k}{n})a_n c_{n-k} + a_n, \quad n = 1, \ldots, P \end{cases} \tag{5}$$

where $P \leq N$ is the desired number of cepstral coefficients.

- Perceptual Linear Prediction analysis (PLP) is a variation of the original linear prediction analysis. The main idea of this technique is to take advantage of three main psychoacoustical properties of the human ear for estimating the audible spectrum, namely: Spectral resolution of the critical band, Equal-loudness curve and Intensity-loudness power law. PLP maps the linear prediction spectrum to the nonlinear frequency scale of the human ear. The perceptual linear prediction coefficients (PLPCC) are an extension of the LPCCs.

4) Robust features:

- In its original formulation, PLPCCs are not robust to signal distorsion. By employing a RASTA (Relative Spectral) filter [29], PLP analysis becomes more robust to distortion. The resulting technique is called RASTA_PLP analysis which consists in a special filtering of the different frequency channels of a PLP analyzer. The RASTA method replaces the conventional critical-band short-term spectrum in PLP and introduces a less sensitive spectral estimation. RASTA_PLP makes PLP more robust to linear spectral distortions.
- Wavelet-based features: We propose a new set of wavelet-based feature vectors, derived from wavelet coefficients. The wavelet coefficients capture time and frequency localized information about the sound waveform that standard Fourier analysis cannot capture: Different from Fourier-based analysis, Wavelet Transforms (WT) use short duration basis functions to measure the signal high frequency content and long duration basis functions for low frequency content (constant-Q analysis) [30].

The WT implements a bank of filters, which are scaled versions of a prototype filter $\psi(t)$ given by:

$$\psi_{\tau,a}(t) = a^{-\frac{1}{2}}\psi(t - \frac{\tau}{a}), \ \ \psi \in L^2 \ (L^2 \text{ is a finite energy space}) \tag{6}$$

where parameters τ and a are called translation and scaling parameters respectively. The term $a^{-1/2}$ is used for energy normalization. The DWT of a signal $\mathbf{s}$ can be obtained as:

$$D(j,k) = 2^{-j/2} \sum_i \mathbf{s}(i)\psi^*(2^{-j}i - k) \tag{7}$$

where i, j and k are integers. By choosing the scaling factor as dyadic (2^j) the resultant transform is known as dyadic DWT. If the wavelet decomposition is computed up to a scale 2^j, the resultant signal representation is not complete. The lower frequencies corresponding to scales larger than 2^j must also be computed and added. These lower frequency components can be evaluated by using the following equation:

$$A(j,k) = 2^{-j/2} \sum_i \mathbf{s}(i)\phi^*(2^{-j}i - k) \tag{8}$$

where $\phi^*(i)$ is the complex conjugate of the scaling function $\phi(i)$.

The DWT can be viewed as the process of filtering the signal using a low pass (scaling) filter and high pass (wavelet) filter. Thus, the first layer of the DWT decomposition of a signal splits it into two bands giving a low pass version and a high pass version of the signal. The low pass signal gives the approximate representation of the signal while the high pass filtered signal gives the details or high frequency variations. The second level of decomposition is performed on the low pass signal obtained from the first level of decomposition. Thus a wavelet decomposition results in a binary tree like structure which is left recursive (where the left child usually represents the lower frequency band). For more information about the Wavelet Transform (WT) and its implementations, interested readers may refer to [31], [30].

In practice, features based on standard sliding window Fourier analysis implicitly assume that the analyzed signal is stationary over each window. However, in the case of surveillance sound signals, impulsive events are often met, and the local stationarity assumption is inaccurate. One could select a short duration window, but this makes the frequency resolution poor, due to the Heisenberg-Gabor inequality [32].

By using wavelets, which somehow implements multiple time resolution analysis, two very short bursts can be separated in time by going to the high frequencies.

Therefore, this analysis can be used for the signals which have short-duration high-frequency components and long-duration low-frequency components. This is indeed a property of most impulsive sounds.

5) Feature combinations: In order to approach the perfect feature set as much as possible, we propose to combine the above features by including them, or not, into the feature vector used to perform classification. This optimisation approach is important essentially in the presence of acoustical variability due to real world background noises. Thus, a set of robust features was selected (PLP_RASTA and Wavelet-based features) for the feature combinations.

Our approach is different from the commonly used feature selection process which consists of choosing a maximal informative subset from a given set of features. Statistical methods, such as the Principal Component Analysis (PCA) [33] that maximizes the variance among the features, are often applied for feature selection. Besides, PCA can be used to generate new features based on existing ones [34]. In this paper, the proposed method is not dealing with features selection but with features vectors selection. It consists of evaluating several features vectors separately and selecting the most performing ones for use as basis features vectors which are necessary issued from basic signal processing transforms (such as DFT, DCT and DWT). Then, to each selected basis vector, another set of vectors will be added and the performance of the final features vector (obtained by concatenating the basis and the added vectors) will be evaluated.

In order to facilitate the study of the composite vector quality, the basis for the feature combination are only the individual features issued from cepstral-based transforms and wavelet-based transforms. This, since, it is noteworthy that individual temporal-based features (Zero-Crossing Rate and short-time average energy) and frequency-based features (Spectral Centroid and Spectral Roll-off point) can not capture critical information for distinguishing between the different sounds, and thus they only perform well if used as additive features.

The research for an optimal solution is achieved by the following strategy. Starting from a well performing features vector we add other features that showed to be independent (i.e. with low redundancy and minimal correlation). Concatenating several basis features vectors is also possible. The composite vector is obtained by concatenating various basis features representing various representation domains. Features or groups of features that do not improve retrieval quality are removed from the combination. Considering several features from various representation domains is very important. Time-based features are extracted from the signal in time domain. Spectral features are derived after the signal has been transformed using one of the signal processing transforms previously described. Whereas, wavelets-based features capture the time-frequency representation of the signal.

The quality of the vectors with combined features is evaluated with the proposed classification technique based on HMMs with gaussian mixtures.

C. The baseline classifier

The classification of a signal is usually performed in two steps. First, a pre-processor employs signal processing techniques to generate a set of features characterizing the signal to be classified. These features form a feature vector. A decision rule is then utilized by the classifier to assign the pattern to a particular class. During the training phase, the classifier will learn how to discriminate between the various classes. Then, unlabelled patterns can be classified by the system during the test phase. This type of learning is called the *supervised learning*.

In our case, for the classification of environmental noise sources, the class could be screams, explosions, gun shots, glass breaks, etc. During the supervised training phase, class labels identifying the elements of a set of training samples are provided to the system so that it can adjust the parameters of the classifier to obtain optimum performance according to the minimization of the error rate criterion [26], [1]. Once the system has been sufficiently trained for a particular pattern recognition application, its parameters are ”frozen” and the classifier is put into service.

In the literature [13], many classifier structures and training methods have been proposed, based on various computational paradigms such as artificial neural networks, fuzzy logic or statistics. In this paper, we will only consider the statistical paradigm, which has been developed with the powerful tools of statistical decision theory [1]. In the statistical approach, the feature vectors are modelled by random variables, the training of the classifier is viewed as a statistical estimation problem, and the classification itself as a hypotheses testing problem.

In the following, let $\Gamma = \{\lambda_1, \lambda_2, ..., \lambda_N\}$ be the set of *N* possible classe labels and assume that the pattern generation mechanism is state λ_i with a priory probability, or prior, $P(\lambda_i)$. For the noise recognition problem, the state of the pattern generation mechanism λ_i is the nature of the noise source (e.g. screams, glass breaks,...) and $P(\lambda_i)$ is equivalent to the proportions of events caused by the noise source λ_i. The *d*-dimensional feature vector $\mathbf{x}$ is obtained at the output of the pre-processor. The feature vector takes its values in a *d*-dimensional subset $\Omega \subset \mathbb{R}^d$ called features space. To each state of the pattern generation mechanism λ_i corresponds a class of patterns with a *class conditional probability density function* (or *pdf*) $p(\mathbf{x} \setminus \lambda_i)$ describing the distribution of the pattern for that particular state. The goal of pattern recognition is to decide on the state of the pattern generation mechanism based on the observation of the realization of the random variables $\mathbf{x}$. That is, we want a decision rule $\hat{\lambda}$ assigning a pattern vector $\mathbf{x}$ to class $\hat{\lambda}(\mathbf{x}) \in \lambda$ for every possible value of $\mathbf{x}$ in Ω.

A common way to represent a classifier is in terms of a set of discriminant functions $g_i(\mathbf{x}), i = 1, ..., n$. The effect of any decision rule $\hat{\lambda}(\mathbf{x})$ is to divide the feature space into *n* disjoint *decision regions* $\Omega_i, ..., \Omega_n$, separated by decision surfaces. Thus, the decision rule can be written as:

$$\hat{\lambda}(\mathbf{x}) = \lambda_i \quad if \quad \mathbf{x} \in \Omega_i, \tag{9}$$

where

$$\Omega_i = \{x \in \Omega : g_i(x) > g_j(x), \forall j \neq i\}. \tag{10}$$

For more details about decision rules see [1].

D. Discussion: Applicability of HMM's to automatic sounds recognition

All the pre-processing methods convert the original acoustic signal into a sequence of continuous-valued vectors. Which allow, in the training phase, the utilization of continuous HMM's. Many sounds recognition systems presented in the literature classify a sound event based on its spectral characteristics only. The feature vector used for classification is computed either over the entire sound event or from a short frame of the sound signal. In the second case, each short frame extracted from the sound event is classified independently. The performance of both methods suffers from the fact that the temporal evolution of the sound event is not taken into account whereas this temporal evolution contains features that can help the classification.

To improve the classification performance, it seems interesting to use a classifier exploiting time-frequency information instead of spectral information only. In fact, features could be computed and the structure of the sequence of these features could also be analyzed to classify the sound event as a whole rather than on a frame by frame basis [1].

In this paper, because of their transient nature, the considered sounds are well suited to be modelled by left-right HMM's. The structure of a HMM will reflect the structure of the modelled process. Left-right HMM's are particularly well suited to model transient signals which have a particular temporal signature. They are commonly used in speech processing to model words. In the context of sounds recognition, left-right models will be used to model impulsive events such as a glass breaks or screams. However, ergodic HMM's are well suited to model stationary signals.

The type of HMM's that will be used must be selected and their parameters including (number of states, transition probability matrix structure, etc.) must be chosen. In fact, a Hidden Markov Model (HMM) consists of a series of states connected together to form a Markov Chain [26]. These states can be desired by an emission probability which is typically a probability density function, and a transition probability matrix. The transition probability matrix will describe the likelihood of the next set of input vectors resulting in a match for the same state or a match in another state.

E. Feature selection procedure

For a sufficiently large training set, it is possible to select the optimal set by applying an original cross-validation procedure [35], as described in Algo. 1.

Algo. 1: Optimization procedure for feature combinations

- **Step 0**: Initialization
  - Select an initial combination of features
- **Step 1**: Iterations
  - Add or remove one feature from the combination, resulting in a new set of features
  - for $p = 1, \ldots, P$, do
    - For $i = 1, \ldots, N$, randomly split $\mathcal{S}_i$ in two approximately equal parts denoted $\mathcal{S}_i^1$ and $\mathcal{S}_i^2$ respectively.
    - Compute the class-conditional *pdf*'s for the sets $\{\mathcal{S}_1^1, \ldots, \mathcal{S}_N^1\}$
    - For each datum **x** in $\{\mathcal{S}_1^2, \ldots, \mathcal{S}_N^2\}$, compute $p(\mathbf{x} \setminus \lambda_i)$ for $i = 1, \ldots, N$ and assign **x** to the class that verifies Eq. (9).
    - Evaluate $\mathcal{E}_p$, the number of misclassifications (i.e., the number of data assigned to a class they do not belong to)
  - compute the average number of errors $\overline{\mathcal{E}} = \sum_{p=1}^{P} \mathcal{E}_p \ / P$

III. Environmental Adaptation of HMMs to Background Noises

To design an automatic sounds recognition system adapted to a new environmental situation, the class-conditional *pdf*'s from the library of pre-trained elements need to be adapted to a new situation. This can be done by the adaptation approach that describes a transformation function *g(.)* for each of the selected classes λ_i. The development of a classifier adapted to a specific background noise can thus be performed as follows. First, we pick the adequate elements from the library of pre-trained source types. Then, we modify their *pdf*'s[5] so that they are suited to the new environment. If there are *N* possible sound sources, each can have its own model of variation $g_i(\mathbf{x})$. Finally, we construct the decision rule based on the modified *pdf*'s.

A. Adaptableness of the classifier to different situations

An automatic sounds recognition system may encounter many different types of sounds and many different measurement conditions[6]. Of course, it is possible to perform a new training of the classifier each time the background noise change. But a new training requires new labelled training samples, which must be provided by a human expert. For that reason, this solution is not acceptable. On the other hand, training a classifier for a large number of environmental conditions is also theoretically possible, but such training is not practical[7]. However, insensitivity to the variations in observation conditions will often be obtained to the detriment of its performance on specific cases. In fact, even with good environment independent (EI) systems some environments are modelled poorly, and it is still the case that environment dependent (ED) systems can give significantly better performance with sufficient environment-specific training data. Possible solutions [1] to preserve the performance when environmental conditions change are adaptable classifiers and adaptive classifiers. Adaptable classifiers are trained in a particular situation but can be adapted by the user to a different situation by 'tuning' some pre-processor or classifier parameters without having to retrain the system completely. However, adaptive classifiers can perform this adaptation automatically

[5] i.e., their means and covariance matrices for Gaussian models

[6] only real world background noises will be considered in our work

[7] it would require an enormous amount of training data that represents the variability of the patterns of all possible classes of sounds

without the need for external supervision. In this paper, some algorithms for the conception of an adaptable classifier are introduced.

B. The multi-style training

Our system uses a Hidden Markov Model (HMM) framework for classifying a range of different sounds. Its originality resides in the HMMs training mode which consists in using both clean and noisy sets [36]. In fact, two training modes can be defined: either training on clean data only or training on clean and noisy (multi-condition) data. The advantage of training on clean data only is the modelling of sounds without distortion by any type of noise. Such models should be suited best to represent all available sound information. The highest performance can be obtained with this type of training in case of testing on clean data only. But these models contain no information about possible distortions. One possible solution is building a library of recognizers for various environmental conditions. The recognizer "closest" to the operating environmental conditions is then picked out of the library. Though training a recognizer for every noisy environment is conceivable, this approach remains time-consuming. Thus a straightforward way to cope with the environment variability is to train an EI system using the approach of multi-style training. This is implemented by pooling data from different acoustical environments, similar to the common strategy for speaker-independent systems, which is to combine training data from a number of speakers.

To achieve environment-independence using multi-style training, data from various acoustical environments will be necessary. The problem is the number of environments. For the similar problem of speaker variability, in [37] the authors found that 80 speakers were needed to achieve speaker independence. It is not clear how many different acoustical environments would be necessary to provide sufficiently broad coverage to obtain environment independence.

Moreover, it was shown in [38] that multi-style training increased the robustness at the expense of sacrificing performance with respect to the case of training and testing on the same condition. In this paper, we propose a multi-style training approach: the training database includes different levels of environmental noises added to the original signals (scenes) and the recognizer can be successfully tested in every noisy environment. Moreover, in order to enhance the system robustness, the proposed solution uses environmental adaptation techniques in the multi-condition training system.

C. Adaptation techniques

1) MLLR and Regression Classes: MLLR was originally developed for speaker adaptation [39] but can equally be applied to situations of environmental mismatch. In MLLR adaptation an initial set of environment independent models are adapted to the new environment by transforming the mean and/or the variance parameters of the models with a set of linear transforms. The transformations are trained so as to maximize the likelihood of the adaptation data with the transformed model set. Originally transformations were estimated only for the mean parameters but recently the approach has been extended so that the Gaussian variances can also be updated [40]. In our work, due to computational reasons, MLLR is only implemented for diagonal covariance, single stream and continuous density HMMs. The transformation matrix used to give a new estimate of the adapted mean is given by

$$\hat{\alpha} = \Omega \ \phi \tag{11}$$

where Ω is $d \times d$ transformation matrix (d is the dimension of the data) and ϕ is the mean vector, $\phi = [\alpha_1 \ \alpha_2 \ ... \ \alpha_d]^T$. The transformation matrix Ω is obtained by solving a maximization problem using the Expectation-Maximization (EM) technique [41]. This technique is also used to compute the variance transformation matrix [40].

The adaptation method based of MLLR can be applied in a very flexible manner, depending on the available amount of adaptation data. If a small amount of data is available then a global adaptation transform can be generated and applied to every Gaussian component in the model set. However, as more adaptation data becomes available, improved adaptation is possible by increasing the number of transformations. Each transformation is now more specific and applied to certain groupings of Gaussian components. Rather than specifying static component groupings or classes, a robust and dynamic method is used for the construction of further transformations as more adaptation data becomes available. MLLR makes use of a regression class tree to group the Gaussians in the model set, so that the set of transformations to be estimated can be chosen according to the amount and type of adaptation data that is available. The tying of each transformation across a number of mixture components makes it possible to adapt distributions for which there were no observations at all. With this process all models can be adapted and the adaptation process is dynamically refined when more adaptation data becomes available.

2) Model Adaptation using MAP: Model adaptation can also be accomplished using a maximum a posteriori (MAP) approach [42]. This adaptation process is sometimes referred to as Bayesian adaptation. MAP adaptation involves the use of prior knowledge about the model parameter distribution. Hence, if we know what the parameters of the model are likely to be (before observing any adaptation data) using the prior knowledge, we might well be able to make good use of the limited adaptation data, to obtain a decent MAP estimate. This type of prior is often termed an informative prior.

For MAP adaptation purposes and in our case, the informative priors used are the environment independent model parameters. The update formula for a single stream for state and a mixture component is

$$\hat{\alpha} = \frac{N}{N+\varepsilon} \ \bar{\alpha} + \frac{\varepsilon}{N+\varepsilon} \ \alpha \tag{12}$$

where ε is a weighting of the a priori knowledge to the adaptation data, N is the occupation likelihood of the adaptation data, α is the environment independent mean and $\bar{\alpha}$ is the mean of the observed adaptation data [42].

3) Model Adaptation using MAP/MLLR: MAP adaptation is specifically defined at the component level and it requires more adaptation data to be effective when compared to MLLR. When larger amounts of adaptation training data become available, MAP begins to perform better than MLLR, due to this detailed update of each component. In fact the two adaptation processes can be combined to improve performance still further, by using the MLLR transformed means as the priors for MAP adaptation (by replacing in Eq. (12) with the transformed mean of Eq. (11)). In this case components that have a low occupation likelihood in the adaptation data, (and hence would not change much using MAP alone) have been adapted using a regression class transform in MLLR.

IV. Experimental set-up and evaluations

A. Database description

The major part of the impulsive sound samples used in the recognition experiments is taken from different sound libraries available on the market [24], [25]. Considering several sound libraries is necessary for building a representative, large, and sufficiently diversified database. Some particular classes of sounds have been built or completed with hand-recorded signals. All signals in the database have a 16 bits resolution and are sampled at 44100 Hz. In this way, all possible audio spectrum components can be exploited for recognition purposes. This point is very important for impulsive sounds, whose frequency bandwidth can be rather extended, because of sharp temporal attacks (such as guns and explosions). Furthermore, some considered sounds show an important energy content in the highest frequencies, as glass breaks for example.

During database construction, great care was devoted to the selection of the signals. When a rather general use of the recognition system is required, some kind of intra-class diversity in the signal properties should be integrated in the database. Even if it would be better for a given recognition system, to be designed for the specific type of encountered signals, it was decided in this study, to incorporate sufficiently diverse signals in the same category. As a result, one class of signals can be composed of very different temporal or spectral characteristics, amplitude levels, and duration and time location.

The selected impulsive sounds belong to the classes listed below. As we can see, all categories are typical of surveillance. The number of considered samples for each sound category is indicated in Table I.

TABLE I
Database description

Classes	Train	Adaptation	Test	Total
human screams	35	12	26	73
explosions	30	11	21	62
glass breaks	40	15	32	88
phone rings	25	11	15	51
door slams	200	33	81	314
dog barks	25	9	21	55
gunshots	150	25	50	225
children voices	40	17	30	87
machines	30	12	18	60
Total	575	145	295	1015

Furthermore, other non-impulsive classes of sounds (machines, children voices) are also integrated in the experimentation. Their utilities come on the occasion of robustness evaluation. We note that the number of items in each class is deliberately not equal, and sometimes very different. Moreover, explosion and gunshot sounds are very close to each other. Even for a person, it is sometimes not obvious to discriminate between them. They are intentionally differentiated, to test ability of the system in separating very close classes of sounds.

B. The Environment Independent (EI) system

The baseline recognizer is trained using the previously described database. One real world background noise[8] is added to each scene at Signal to Noise Ratios (SNRs) ranged from -10dB to 30dB. The obtained environmental independent system is called the baseline recognizer and is trained on more than 5500 different sounds. In real applications, the EI system will be able to classify various sounds recorded in different environmental conditions. To make preliminary experiments in order to test that system, it is necessary to choose a particular environment (for example a real world background noise at SNR=5 dB) and to adapt the EI models by adaptation data which are contaminated by the same noise at the same level. Later we will demonstrate that the obtained adapted system performs the baseline recognizer.

C. The adapted system parameters

Our objective is to adapt the current well trained environment independent models to the characteristics of a particular environment using a small amount of adaptation data. Thus, we trained acoustic models on artificially perturbed sounds material. Experimental evaluation on environmental adaptation using MAP, MLLR and MAP/MLLR techniques illustrates a recognition improvement over the baseline system results. We assume that for MLLR we update the mean and/or the variance model parameters which are designed in the figure 1 by (m) and (v). The three algorithms are applied on the database previously described for supervised adaptation experiments using various amounts of adaptation data. The smallest set contains 10 scenes and the largest one contains 156 scenes. The efficiency of these adaptation formulations is visible in Fig. 1 (a).

In the experiments, features are computed from all the samples in each sound. The analysis window is Hamming with length 25 ms, with 50 % overlap. The feature vectors are formed with MFCCs. Evaluations are obtained using a HMM-based classifier with 3 hidden states (left-to-right) and 5 Gaussians components in each state (whose parameters are learnt by 3 iterations in the Baum-Welch algorithm [43], [41]).

As we can see, all the adaptation methods lead to an important improvement of the recognition accuracy when compared with the result done by the baseline system (without adaptation). By using 156 adaptation scenes, the error rate of the adapted recognizer improves over the baseline system by more than 30% for rather all the considered algorithms.

[8]This noise was extracted from the Noisex-92 database and re-sampled at 44100 Hz.

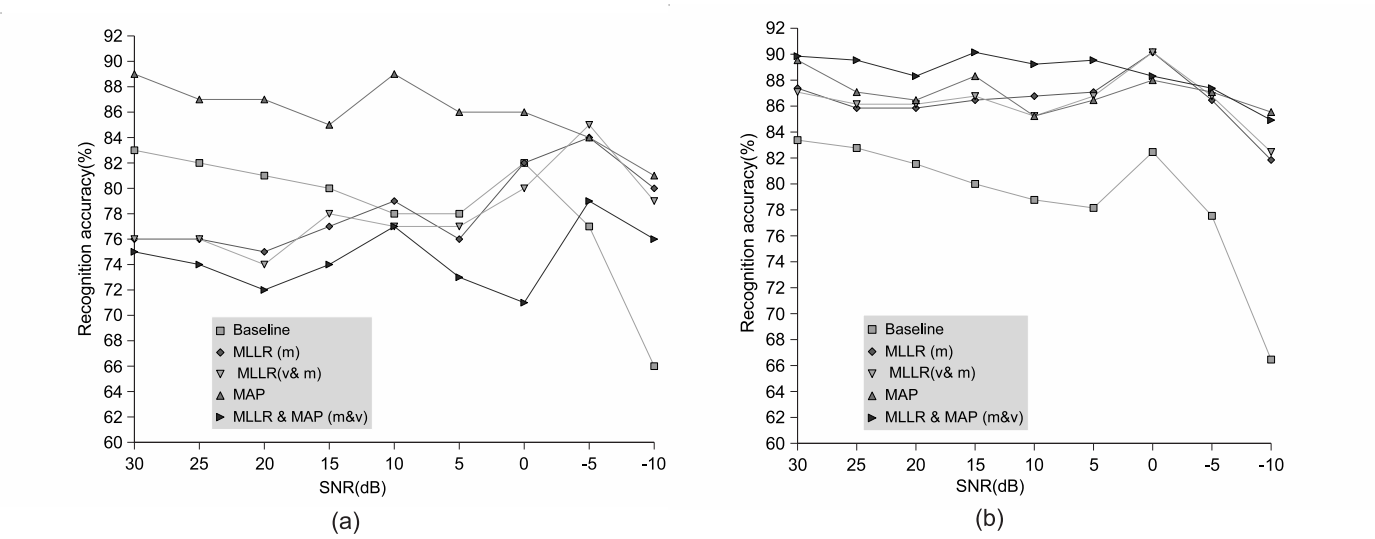

Fig. 1. (a) Recognition rates (%) by applying different adaptation algorithms for various SNR (dB) values (10 utterances are used), (b) Recognition rates (%) by applying different adaptation algorithms for various SNR (dB) values (156 utterances are used)

The efficiency of the MAP/MLLR formulation is then illustrated for a variable amount of adaptation data where an additional improvement is obtained over the MLLR and MAP results. For only 50 scenes and in bad operating conditions (SNR=5 dB) the recognition rate increases from 78.15% to 90.15% (Table II).

TABLE II
RECOGNITION ACCURACY (%) FOR VARIOUS AMOUNTS OF ADAPTATION DATA (N) BY USING VARIOUS ADAPTATION METHODS

Real world background noise (RSB=5 dB)				
N	MLLR(m)	MLLR(m&v)	MAP	MAP/MLLR(m&v)
10	76	77.23	86.46	73.85
30	84	84.00	88.00	89.23
50	87.08	85.54	88.62	90.15
70	88.31	87.38	89.23	89.95
100	86.77	86.15	86.46	89.23
130	87.69	87.08	87.69	89.54
156	87.08	86.77	86.46	89.54

In order to have a rapid adaptation, many tests are done using a small amount of adaptation data. Fig. 1 (b) shows that 10 scenes are sufficient to have good improvement of the recognition accuracy by applying the MAP algorithm.

D. Results with individual features

We have presented above classification results obtained when retaining only a feature vector based on MFCCs. Then, in order to enlight the usefulness of the features presented in Subsection II-B we will present here the results done using the other individual features.

Tab. III provides the results obtained by the adaptation techniques, where the performance rate is computed as the percentage number of sounds correctly recognized and it is given by $(\mathrm{H}/\mathbf{N}) \times 100\%$, where H is the number of correct sounds and $\mathbf{N}$ is the total number of sounds to be recognized.

The use of wavelet coefficients is motivated by their ability to capture important time and frequency features. The Daubechies wavelets with 4 vanishing moments are used.

RASTA_PLP is an improvement of the traditional PLP method and consists in a special filtering of the different frequency channels of a PLP analyzer. The RASTA method replaces the conventional critical-band short-term spectrum in PLP and introduces a less sensitive spectral estimation. RASTA_PLP makes PLP more robust to linear spectral distortions.

E. Effects of features combination

In this section, we illustrate that different feature combinations can lead to quite different performance, see Tab. IV, where the combinations displayed are obtained by including incrementally in the combinations other features that are known to be little correlated with the already selected ones (feature correlations has been investigated in many previous works, see [13], [44] for instance). The results for each combination are evaluated by the HMM-based classifier.

In [44], it is shown that the information of LPC coefficients is already captured by the more expressive MFCCs. Our experiments confirm this conclusion. This is also true for PLP features. Hence, we do not include LPC and PLP coefficients in the combination whenever MFCCs are already included. For highly redundant groups of features we choose only individual representative components. [13] shows that adding temporal features can improve the classification performance. Thus, we added ZCR and the average energy which are one-dimensional features. ZCR is closely related to the fundamental frequency of frame. In the case of environmental sounds the fundamental frequency may be similar for different classes; this is why ZCR is not suited to classification as a single feature. Due to the low-dimension of the tested temporal features (ZCR and the average energy) and the frequency features (SRF and SC), they fail to represent data information. Nevertheless, these features may improve retrieval quality in combination with the preselected basis features (as showed in Subsection II-B).

TABLE III
EFFECTS OF ADAPTATION (156 ADAPTATION SCENES) USING VARIOUS INDIVIDUAL FEATURES.

Features	Non Adapted System		Adapted System		
	Clean Data SNR=40dB	Noisy Data SNR=5dB	MAP SNR=5dB	MLLR SNR=5dB	MAP/MLLR SNR=5dB
PLP	91.38	77.08	77.08	83.15	88.23
LPC	91.69	76.77	80.23	83.15	88.54
$MFCC$	92.9	77.15	84.69	89.02	90.15
DWC	89.46	70.23	77.08	77.08	78.69
$PLP\_ RASTA$	89.23	78.46	84.69	84.69	84.69

TABLE IV
EFFECTS OF FEATURES COMBINATIONS.

Features	Number of features	Recognition Rate (%)
$MFCC + Energy + RF + centroid + ZCR$	16	93.73
$DWC + MFCC + Energy + log\ energy + RF + centroid + ZCR$	67	93.8
$PLP\_ RASTA + Energy + RF + centroid + ZCR$	16	93.2

In general, combinations involving spectral and time-based features are useful, since they combine information of the two complementary domains: while spectral features characterize the frequency contents, time-based features incorporate temporal information and loudness.

It is clear that some features are not able to discriminate between the classes successfully when used alone. The DWT coefficients do not separate wel some classes. This can be partly explained by the fact that DWT coefficients are mostly informative for low frequencies, and they tend to neglect high frequencies. This justifies the use of 12 MFCCs in addition to DWT coefficients. As can be seen in Tab. IV, this significantly improves the discrimination ability.

V. CONCLUSION

In this paper, we have addressed the problem of the automatic recognition of environmental sounds. Since, our work was motivated by a practical surveillance application, it was necessary to be able to classify auditory scenes under important noise degradation conditions.

Thus, it was demonstrated that by adapting the system models and then, by applying a features combinaison method, a visible improvement in the discrimination ability of an automatic sounds recognition system can be showed, even under important noise degradation conditions.

There are many interesting directions in which to continue the research. others classifiers can be studied and more research is needed on studying how to select the best features combinations for each type of classifier.

REFERENCES

[1] C. Couvreur, "Environmental Sound Recognition: A Statistical Approach," Ph.D. dissertation, Faculte Polytechnique de Mons, Belgium, June 1997.
[2] V. Peltonen, "Computational auditory scene recognition," Ph.D. dissertation, Tampere University of Technology, Finland, 2001.
[3] D. Istrate, "Détection et reconnaissance des sons pour la surveillance médicale," Ph.D. dissertation, INPG, France, Dec. 2003.
[4] K. El-Maleh, "Frame level noise classification in mobile environments," Ph.D. dissertation, McGill University, Montreal, Canada, Jan. 2004.
[5] R. S. Goldhor, "Recognition of environmental sounds," in *ICASSP*, vol. 1, New York, USA, 1993, pp. 149–152.
[6] B. Uvacek, H. Ye, and G. Moschytz, "A new strategy for tactile hearing aids: tactile identification of preclassified signals (tips)," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, New-York, USA, May 1988.
[7] A. K. S. Oberle, "Recognition of acoustical alarm signals for the profoundly deaf using hidden markov models," in *International Symposium on Circuits and Systems*, vol. 1, Seattle, USA, 1995, pp. 2285–2288.
[8] J. A. Osuna and G. S. Moschytz, "Recognition of acoustical alarm signals with cellular networks," in *European Conference on Circuit Theory and Design*, Istanbul, Turkey, 1995.
[9] M. J. Paradie and S. Nawab, "Classification of ringing sounds," in *ICASSP*, Apr. 1990.
[10] R. H. Cabell, C. Fuller, and W. O'Brien, "Identification of Helicopter noise Using a Neural Network," *AIAA Journal*, vol. 30, no. 3, pp. 624–630, Mar. 1992.
[11] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in *ICASSP*, Istanbul, Turkey, 2000, pp. 753–756.
[12] H. Soltau, T. Schultz, and M. Westphal, "Recognition of music types," in *ICASSP*, Seattle, WA, 1998.
[13] A. Dufaux, "Detection and recognition of Impulsive Sounds Signals," Ph.D. dissertation, Faculté des sciences de l'Université de Neuchâtel, Switzerland, 2001.
[14] A. Bregman, *Auditory scene analysis*. Cambridge, USA: MIT Press, 1990.
[15] K. D. Martin, "Sound-source recognition: A theory and computational model," Ph.D. dissertation, MIT Press, 1999.
[16] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*. New York: Springer, 2006.
[17] M. Orr, D. Pham, B. Lithgow, and R. Mahony, "Speech perception based algorithm for the separation of overlapping speech signal," in *The Seventh Australian and New Zealand Intelligent Information Systems Conference*, 2001.
[18] M. Cowling, "Non-speech environmental sound classification system for autonomous surveillance," Ph.D. dissertation, Faculty of Engineering and Information Technology, Griffith University, 2004.
[19] M. Cowling and R. Sitte, "Recognition of environmental sounds using speech recognition techniques," *Advanced Signal Processing for Communications Systems*, 2002.
[20] ——, "Comparison of techniques for environmental sound recognition," *Pattern Recognition Letters*, vol. 24, pp. 2895–2907, 2003.
[21] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261–291, 1995.
[22] C. H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Communication*, vol. 25, pp. 29–47, 1998.
[23] ——, "Adaptive classification and decision strategies for robust speech recognition," in *Workshop on Robust Methods Speech Recognition Adverse Conditions*, Tempere, Finland, May 1999.
[24] Real World Computing Paternship, "Cd-sound scene database in real acoustical environments," http://tosa.mri.co.jp/sounddb/indexe.htm, 2000.
[25] Leonardo Software, Santa Monica, USA, http://www.leonardosoft.com.

[26] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. of IEEE*, vol. 77, no. 2, pp. 257–289, Feb. 1989.
[27] P. Mermelstein and S. B. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *ICASSP*, vol. 28, 1980, pp. 357–366.
[28] J. Makhoul, "Linear prediction: A tutorial review," in *Proceedings of IEEE*, vol. 63, 1975, pp. 561–580.
[29] P. Mermelstein and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, 1994.
[30] M. Vetterli and J. Kovacevic, *Wavelets and subband coding*. Englewood Cliffs, NJ, USA: Prentice Hall, 1995.
[31] S. Mallat, *A wavelet tour of signal processing*. Academic Press, 1998.
[32] P. Flandrin, *Time-frequency/time Scale Analysis*. San Diego, USA: Academic Press, 1999.
[33] I. Jollife, *Principal Component Analysis*. New York, USA: Springer-Verlag, 1986.
[34] J. Loehlin, *Latent variable models: An Introduction to Factor, Path, and Structural Analysis*. Lawrence Erlbaum Assoc., 2001.
[35] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New-York, USA: Springer, 2001.
[36] A. Rabaoui, Z. Lachiri, and N. Ellouze, "Hidden Markov model environment adaptation for noisy sounds in a supervised recognition system," in *International Symposium on Communication, Control and Signal Processing (ISCCSP)*, Marrakech, Morroco, Mar. 2006.
[37] K. Lee and H. Hon, "Large-vocabulary speaker-independent continuous speech recognition," in *ICASSP*, Apr. 1988.
[38] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," Ph.D. dissertation, Department of Electrical and Computer Engineering, Carnegie Mellon University, 1990.
[39] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.
[40] M. J. F. Gales and P. C. Woodland, "Variance compensation within the mllr framework," Technical Report CUED, Cambridge University, Tech. Rep., 1996.
[41] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," International Computer Science Institute, Berkeley, USA, Tech. Rep., 1998.
[42] K. Shinoda and C.-H.Lee, "Unsupervised adaptation using structural bayes approach," in *ICASSP*, 1998.
[43] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.
[44] D. Mitrovic, "Discrimination and Retrieval of Environmental sounds," Ph.D. dissertation, Vienna University of Technology, Dec. 2005.