PAN ARCHIVES OF ACOUSTICS
Vol. **37**, No. 4, pp. 423–434 (2012)

DOI: 10.2478/v10168-012-0054-z

Hierarchical Classification of Environmental Noise Sources Considering the Acoustic Signature of Vehicle Pass-Bys

Xavier VALERO, Francesc ALÍAS

GTM – Grup de Recerca en Tecnologies Mèdia, La Salle – Universitat Ramon Llull
Quatre Camins 30, 08022 Barcelona, Catalonia, Spain; e-mail: {xvalero, falias}@salle.url.edu

(*received March 29, 2012; accepted August 7, 2012*)

This work is focused on the automatic recognition of environmental noise sources that affect humans' health and quality of life, namely industrial, aircraft, railway and road traffic. However, the recognition of the latter, which have the largest influence on citizens' daily lives, is still an open issue. Therefore, although considering all the aforementioned noise sources, this paper especially focuses on improving the recognition of road noise events by taking advantage of the perceived noise differences along the road vehicle pass-by (which may be divided into different phases: approaching, passing and receding). To that effect, a hierarchical classification scheme that considers these phases independently has been implemented. The proposed classification scheme yields an averaged classification accuracy of 92.5%, which is, in absolute terms, 3% higher than the baseline (a traditional flat classification scheme without hierarchical structure). In particular, it outperforms the baseline in the classification of light and heavy vehicles, yielding a classification accuracy 7% and 4% higher, respectively. Finally, listening tests are performed to compare the system performance with human recognition ability. The results reveal that, although an expert human listener can achieve higher recognition accuracy than the proposed system, the latter outperforms the non-trained listener in 10% in average.

Keywords: acoustic signature, environmental noise monitoring, Gaussian Mixture Models, hierarchical classification, Mel Frequency Cepstral Coefficients, sound classification, traffic noise, vehicle pass-by.

1. Introduction

Environmental noise might be regarded as unwanted sound produced by transport, industrial or recreational activities (EU Directive, 2002). Those environmental noise sources typically encountered on cities and urban areas affect citizens' quality of life, besides involving harmful health effects (Babisch, 2006; Rasche, 2004). The publication and adoption of the Green Paper on Future Noise Policy in Europe in 1996 (EU Commission, 1996) contributed to the awareness of environmental noise as a pollutant. Six years later, the Environmental Noise Directive (END) (EU Directive, 2002) was published with the latest goal of informing the public about their exposure to noise and drawing up appropriate action plans to prevent the harmful effects derived from their exposition to noise. In compliance with the END, the member states of the European Union are required to report the noise levels caused by the aforementioned sources by producing strategic noise maps in their main cities, transport infrastructures and industrial sites. In the same context, the last review of the International Standard on the Determination of Environmental Noise Levels (ISO 1996-2:2007) states that in traffic noise assessment, vehicles have to be classified within at least two categories: light and heavy (ISO, 2007). In this framework, measurements in complex acoustic situations (e.g., urban environments) have to be conducted, with the presence of noise from diverse origins, such as road traffic, railway traffic, aircrafts, industrial, etc. Only if we are capable of developing precise descriptions and measurements, the action plans designed to reduce or prevent high levels of environmental noise will be efficiently addressed.

In this sense, the implementation of environmental noise recognition systems may provide with an automatic transcription of the types of noise sources present on a certain location and their contribution to the overall noise level measured. The application of those systems would be of special interest for long-term measurements (lasting from several hours up to

several months), as a means for recreating the acoustical situation automatically. Typically, the environmental noise recognition systems are composed of two main blocks: the first one consists in signal processing, which parameterises the sound signals by computing a set of representative features, whilst the second step performs the recognition of the environmental noise events by means of some machine learning technique, generally following a supervised learning approach.

In this work, we consider the environmental noises that need to be mapped according to the END, i.e., railway, air transport, industrial and road traffic noise. As a first step, we address the classification of the aforementioned noise sources, leaving their detection and their identification among noise mixtures for future works. It should also be noted that, in order to enable embedding the technology into classical noise monitoring stations (typically composed of one sound level meter), we are interested in computationally efficient solutions besides discarding approaches based on different microphones (Mato–Méndez, Sobreira–Seoane, 2011).

The first contribution of the paper is to extend and update the comparison of signal features and machine learning techniques for the problem at hand with respect to (Valero, Alías, 2011b) by considering up to 13 signal features and 4 machine learning techniques, thus yielding a collection of 52 tested combinations. Nevertheless, the main contribution of the paper resides in the classification of road vehicle noise sources. The reason to especially address this problem is twofold: firstly, because it is the type of environmental noise with largest impact on the citizens' quality of life; and secondly, because the different vehicles (cars, trucks and scooters) present very similar acoustic signatures, which makes them the most difficult noise sources to be distinguished, according to the results reported in previous works (Defréville *et al.*, 2006; Ntalampiras *et al.*, 2008; Valero, Alías, 2011b). Therefore, the main aim of this study is to improve the classification of such noise sources by taking advantage of the change in the perceived sound along the pass-by of the vehicle from the receiver position as the basis for the discrimination among road vehicle noise sources.

The rest of the paper is organized as follows. Section 2 introduces the related work in environmental noise source recognition. Section 3 reviews the background in sound signal features and machine learning techniques. Section 4 describes the characteristics of the road noise sources, which is the basis of the proposed classification scheme. Section 5 describes the experimental evaluation and Sec. 6 details the analysis of the obtained results. Section 7 discusses the results and, finally, Sec. 8 draws up the conclusions and future work.

2. Related work

Up to our knowledge, one of the first approaches focused on recognising environmental noise events recorded at noise monitoring stations was presented in (Couvreur *et al.*, 1998). The recognition system was borrowed from the speech recognition field, including both Linear Predictive Coefficients (LPC) and Hidden Markov Models (HMM) so as to classify five different types of noise sources: cars, trucks, mopeds, aircrafts and trains. In (Rabaoui *et al.*, 2004), several signal features were explored: LPC, Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Predictive (PLP), Discrete Wavelet Coefficients (DWC) and Mel Frequency Discrete Wavelet Coefficients (MFDWC). The performance of these features was experimentally compared using HMM on a corpus composed of five noise events (cars, trucks, planes, trains and dogs). Among them, PLP and MFCC attained the best classification results. A broader signal feature comparison was carried out in (Valero, Alías, 2011b), extending to 13 the number of considered signal features (including temporal domain, spectral domain, linear prediction and Wavelet features). In combination with a Multilayer Neural Network, both MPEG-7 and MFCC attained the highest averaged recognition accuracies in a corpus containing road vehicles, aircrafts, trains and industrial noises. Besides comparing different signal features, Fisher Linear Discriminant and K-Nearest Neighbor (KNN) were evaluated in (Sobreira *et al.*, 2008) for the recognition of specifically road traffic noise sources (cars, trucks and scooters). Experimental results showed that KNN was the machine learning technique yielding the best performance, specifically when considering feature combination (MFCC, Sub-band Energy Ratio (SBER) and Spectral Roll-Off (SRO)).

Broadly speaking, the different aforementioned research works were reduced to experiment with different signal features and machine learning techniques following a flat classification scheme. In contrast, (Defréville *et al.*, 2006) and (Ntalampiras *et al.*, 2008) proposed addressing the problem by means of a hierarchical scheme. In (Defréville *et al.*, 2006), the classification system was composed of a first layer that discriminated between mechanic (moped, bus, motorcycle and car) and non-mechanic classes (birds and voices). The system evaluation consisted in conducting many one-against-all experiments, and no test was carried out involving the six urban sounds altogether. In addition, for every one-against-all experiment a different combination of features (from a library of 80 operators such as FFT, min, max, arcsin, etc.) was selected. Thus, both the global performance of the system and its generalization capability still remain unknown. Similarly, in (Ntalampiras *et al.*, 2008) a hierarchical classification scheme was proposed, com-

posed of a first Gaussian Mixture Model (GMM) that classified the samples into mechanic and non-mechanic categories, followed by HMM that completed the classification of the specific noise source within each category. This work included experiments considering all the environmental noises at the same time, thus providing information about the performance and the most frequent class confusions of the system when trying to recognize all the noise sources at the same time. Both with MFCC and MPEG-7 features, the most common misclassifications were observed among the mechanic classes, especially observable in the car category.

In this work, we make a significant step further from previous hierarchical classifiers approaches in order to specifically improve the recognition of road vehicle noise sources (while also considering aircrafts, trains and industrial noise). Hence, we put forward an environmental noise classification scheme that takes into account the particular acoustic signatures of road vehicle noise sources by dividing the vehicle pass-by into different phases (see Sec. 4).

3. Background review

In this section we briefly review the main signal features and machine learning techniques employed in the related literature, which are later considered in the experiments described in Sec. 5.

3.1. Signal features

One of the main goals the sound signal features may accomplish is that they should accurately represent the characteristics of the sound signals by a reduced amount of coefficients. According to the related literature, the choice of the signal feature is particularly important for environmental noise classification (Umapathy *et al.* 2005; Chu *et al.,* 2009). In the past, a wide variety of signal features have been employed to describe sound signals. In this section, we briefly describe the most frequently used. We refer the interested reader to (Kim *et al.* 2005; Rabiner, Juang, 1993; Eronen *et al.* 2006) and (Rabaoui *et al.* 2004) for a more detailed description.

a) Time-domain features:

- Short Term Energy (STE): describes the time envelope of the signal, being calculated as:

$$STE = \sum_{n=0}^{N-1} |x[n]|^2 = \sum_{k=0}^{K-1} |X[k]|^2, \qquad (1)$$

where $x[n]$ states for the sound signal in the time domain, $X[k]$ is its Fourier Transform, N is the number of samples of the signal frame analysed and K the number of FFT points.

- Zero Crossing Rate (ZCR): number of times that the signal crosses the zero in terms of amplitude. It is related to the periodicity of the signal.

b) Spectral domain features:

- Spectral Centroid (SC): measures the centre of gravity of the power spectrum $X[k]$ (2).

$$\mathrm{SC} = \frac{\sum_{k=0}^{K-1} k \cdot |X[k]|}{\sum_{k=0}^{K-1} |X[k]|}. \qquad (2)$$

- Spectral Roll-off (SRO): bandwidth in which is concentrated most of the power spectrum energy (3). It gives information of the "skewness" of the spectral shape.

$$\mathrm{SRO} = \max_m \left(\sum_{k=0}^{m} |X[k]| \leq TH \cdot \sum_{k=0}^{K-1} |X[k]| \right), \quad (3)$$

where TH is set between 0.88 and 0.95 (Kim *et al.*, 2005).

- Sub-Band Energy Ratio (SBER): energy distribution along the sub-bands B_i with respect to the total signal spectrum energy.

$$\mathrm{SBER}_i = \frac{\sum_{k \in B_i} |X[k]|}{\sum_{k=0}^{K-1} |X[k]|}. \qquad (4)$$

- Mel Frequency Cepstral Coefficients (MFCC): Discrete Cosine Transform of a log power signal spectrum on a non-linear mel frequency scale.
- MPEG-7 features: a total of 15 different low-level descriptors are defined in the MPEG-7 standard (ISO, 2001), considering different aspects of the sound signal.
- Spectral Flatness (SF): deviation of the signal power spectrum with respect to a flat spectrum for each of the predefined spectral bands (5).

$$\mathrm{SF}_i = \frac{\sqrt[hi_i - lo_i]{\prod_{k=lo_i}^{hi_i} X[k]}}{\frac{1}{hi_i - lo_i} \sum_{k=lo_i}^{hi_i} X[k]}, \qquad (5)$$

where hi_i and lo_i are respectively the upper and lower band cut-off frequencies.

c) Linear prediction features:

- Linear Predictive Coefficients (LPC): coefficients a_i extracted from the prediction of the current sample as the linear combination of the p previous samples $x(n)$.

$$\widetilde{x}(n) = \sum_{k=1}^{p} [a_k x(n-k)],$$
$$\mathrm{LPC} = \{a_1, a_2, \ldots\ a_k\}. \qquad (6)$$

- Linear Prediction Cepstral Coefficients (LPCC): Cepstrum c_i obtained from LPC coefficients a_i (7), where G is the filter gain and p is the order of the LPC coefficients.

$$c_0 = \log(G),$$
$$c_i = -a_i + \frac{1}{i}\sum_{k=1}^{i-1}[-(i-k)a_k c_{(i-k)}] \quad 1 \leq i \leq p. \tag{7}$$

- Perceptual Linear Prediction (PLP): modified version of LPC that adds critical-band spectral resolution, equal-loudness pre-emphasis and the intensity-loudness power law to mimic human hearing processing.

d) Wavelet analysis:

- Discrete Wavelet Coefficients (DWC): coefficients extracted with the time-frequency Wavelet Transform.
- Mel Frequency Discrete Wavelet Coefficients (MFDWC): modification of the MFCC that uses the Wavelet transform rather than the Discrete Cosine Transform to compress the signal energy.

3.2. Machine learning techniques

The mission of the machine learning technique is to perform the classification of specific noise events, after being parameterized with any of the aforementioned input signal features. Following a supervised learning approach (Jones, 2008), many different learning algorithms can be employed. In the next paragraphs, we describe some of the most typical ones in the context of environmental noise classification. For a deeper discussion, see (Bishop, 2003; Breiman *et al.*, 1993; Rabiner, 1989) and (Jones, 2008).

- Decision Tree (DT): performs the classification according to a set of rules that divide the feature space into several regions. Each rule is associated to one node of the tree and it is based on a single data attribute. Different rules can be applied to split data at the node level (e.g., maximum deviance reduction, Gini's diversity index, etc.). After building the tree, a pruning process is typically performed to avoid a too specific data fitting during the training stage.
- K-Nearest Neighbour (KNN): performs the classification based on a majority vote of the K closest neighbours to the sample being classified. The method shows a good trade-off between simplicity (no training process is required) and accuracy.
- Neural Network (NN): biologically-inspired computational model that provides general parameterised non-linear mappings between a set of input and output variables. The mapping is performed by means of several weighting nodes and activation functions. Depending on their architecture, several NN may be built: from the simplest Perceptron to more complex Multi-Layer NN or recursive NN.
- Gaussian Mixture Model (GMM): models the probability density function of each noise class as the weighted sum of M simple Gaussian functions, where each Gaussian function is represented by the mean and the covariance matrix of the data. Then, Expectation-Maximization algorithm is frequently employed to identify the parameters of each class yielding higher conditional probability.
- Hidden Markov Model (HMM): unlike the aforementioned machine learning techniques, it considers the time evolution of the noise signals by modelling them as a finite sequence of unobservable states, being each one modelled by means of a certain probability distribution.

4. Road vehicle pass-by phases recognition

In this section, we first describe the characteristics of the road vehicle noise sources that motivate the proposed classification scheme, which is subsequently detailed.

4.1. Characteristics of road noise sources

A road vehicle may be modelled as the combination of four noise sources with different characteristics originated by: *i*) the engine, *ii)* the contact between the tire and the asphalt, *iii)* the exhaust pipe, and *iv)* the aerodynamic effect (Cevher *et al.*, 2009). The engine noise contains both a deterministic/harmonic component (caused by the fuel combustion in the cylinders) and a stochastic component (due to the turbulent air flow in the air intake) (Amman, Das, 2001). The tire noise is the main noise source when the speed of the vehicle is higher than 50 km/h, and it is composed of a vibration component (originated by the contact between the tire and the asphalt, with an important spectral content between 100 Hz and 1 kHz) and an air component (originated from the air being sucked in or forced out of the tire, with a dominant spectrum between 1 kHz and 3 kHz) (Sandberg, Ejsmont, 2002). It should also be noted that, in the direction of the car, the road and the tire form a geometrical structure that amplifies the noise generated due to their interaction (Cevher *et al.*, 2009). This phenomenon is called the *horn* effect, and especially amplifies the frequencies in the range from 600 Hz to 2 kHz (Sandberg, Ejsmont, 2002). The exhaust system goes from the engine compartment to the back of the car, generating a noise that, due to the system distribution, is more noticeable in the rear that in the front of the vehicle (Cevher *et al.*, 2009). Finally, the boundary layer of the vehicle generates an air flow which produces an aerodynamic noise quite loud but only at very high speeds (Peeters, Blokland, 2007).

To sum up, the noise generated by a vehicle is characterized by a complex spectral content due to the variety of independent noise but also for their spatial distribution (Peeters, Blokland, 2007), since the independent noise sources are physically located in different positions of the vehicle. The engine noise might usually be located in the front, the tire noise on the sides and the exhaust pipe noise at the rear of the vehicle. Therefore, and considering the acoustic shadow effect caused by the vehicle itself, those three independent noise sources will have different contribution to the overall sound signal perceived by the receiver at each point of the vehicle pass-by.

On the other hand, with regard to the relative position between the moving vehicle and the receiver, the Doppler effect must be taken into account (Cevher *et al.*, 2009). As it is well known, the signal frequency increases in the vehicle approaching phase, and decreases in the receding phase. The frequency f perceived by the receiver can be calculated as follows:

$$f = \left(\frac{c + v_r}{c + v_s}\right) f_0, \tag{8}$$

where c is the propagation speed of the acoustic wave, v_r is the running speed of the receiver, v_s is the running speed of the source and f_0 is the original frequency of the sound source.

Let us take as example a vehicle driven at 80 km/h. At this speed the tire noise is the predominating source, with a notable energy content centred at approximately 1000 Hz (Sandberg, Ejsmont, 2002). Under those conditions, the signal frequency changes from 1239 Hz when the vehicle is approaching, to 939 Hz when the vehicle is receding (calculated from (1)). Hence, it results in a frequency variation of 300 Hz, which should not be neglected in the classification stage.

The multi-source representation of the vehicle (considering both its complex spectral content and its spatial distribution), together with the Doppler effect as it is moving noise source, are the two main causes of the perceived noise change from the point of view of the receiver. In consequence, the acoustic signature of a vehicle pass-by can be divided into three phases, depending on its position with respect to the receiver or measurement point (microphone) when: the vehicle is approximating (*approaching*), the vehicle is perpendicular to the microphone (*passing*), and the vehicle is moving away (*receding*). This effect might be observed in the spectral representation of a scooter pass-by (see Fig. 1), where also the higher impact of exhaust pipe noise in the receding phase can be noticed. The differences between the approaching and the receding phase might also be noticed in the asymmetry on the time envelope of the road vehicle pass-by (see Fig. 2). Our proposed approach for the automatic classification of noise sources, which is further detailed in the next section, has been built so as to make the most of the spectro-temporal differences of vehicles pass-bys.

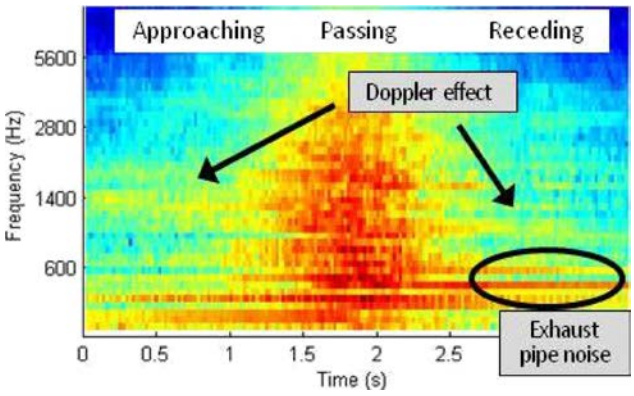

Fig. 1. Spectral differences observed in a scooter pass-by: approaching, passing and receding phases.

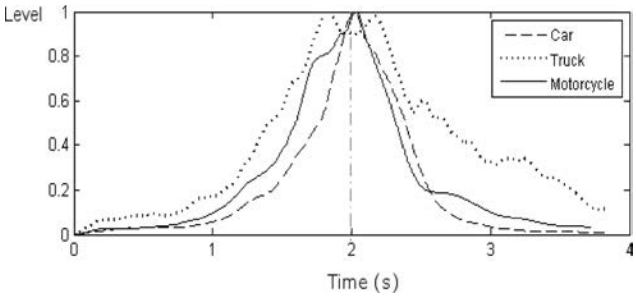

Fig. 2. Signal time envelope of the pass-by of three different road vehicles, after being normalized by its maximum.

4.2. A classification scheme adapted to the spectro-temporal characteristics of road vehicle pass-by

In this section, we describe the classification scheme designed to improve the identification of road vehicle noise sources by considering the characteristics of their acoustic signatures.

As indicated in the introduction, other common environmental noise sources considered in this work (i.e., trains, aircrafts and industry) do not share the same pass-by characteristics of road vehicles. Their acoustic signatures are either continuous (i.e., industry) or present a too long pass-by to allow a short-term response from the recognition system (i.e., an aircraft and a train pass-by can last around 60s and 30s, respectively). Hence, the road vehicles and the remaining noise sources are modelled independently. Taking into account these particularities, we put forward a hierarchical classification scheme to perform the identification of urban noise sources (see Fig. 3).

The process starts by windowing the sound signal and extracting the signal features from each resulting frame. Next, the signal features from all the frames are merged into a single feature vector. In this way, we

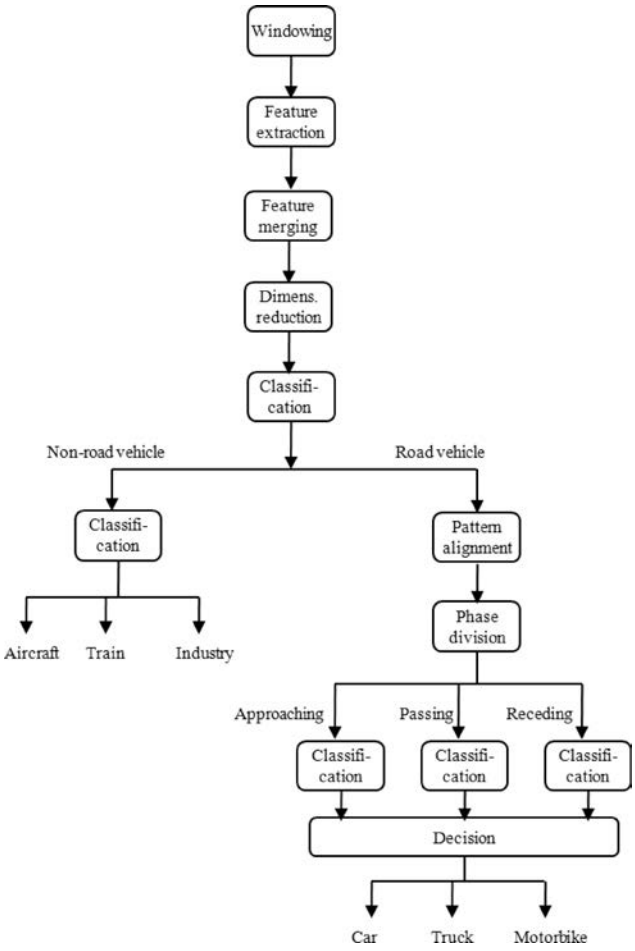

Fig. 3. Block diagram of the proposed hierarchical environmental noise classification scheme adapted to the spectro-temporal characteristics of vehicles pass-by.

manage to preserve the signal time evolution (which is especially relevant in environmental sound recognition (Gygi, 2001)) when employing machine learning techniques that do not necessarily take this evolution into account (i.e., DT, KNN, NN and GMM). As a consequence, the resulting feature vector has a large dimensionality; therefore we next apply PCA to compact the information (Kim *et al.*, 2005). Subsequently, in a first step, the classification scheme divides the noise sources into road vehicle and non-road vehicle categories. In a second step, the specific type of noise source within each broad category is identified. Whilst non-road vehicles are treated following a flat approach, the recognition scheme of road vehicles divides the time pattern of the parameterized noise sample into three parts, corresponding to the approach, passing and receding phases of the vehicle pass-by (see Subsec. 4.1). An independent recognition decision (i.e., class assignment) is taken for each of these three phases and, finally, a second decision layer applies a simple voting scheme to come up with a unique solution (i.e., the recognised noise source). In case of tie (i.e., all three noise sources are labelled as a different type of vehicle), the second decision layer selects the noise source identified at the central passing phase (since it is the phase when the source is closer to the receiver, and thus, we consider it to be less contaminated by background noise).

It is worth noting that the proposed road vehicle classification scheme involves a process to align the centre of the pass-by with the centre of the temporal pattern (see Fig. 4). This process, which is only applied to the road vehicle branch of the hierarchical classification scheme, is carried out by calculating the STE (1) all along signal temporal envelope and detecting the frame with the largest STE value. The sound signal is shifted accordingly to allow the frame with maximum energy becoming the central frame of the sound pattern.

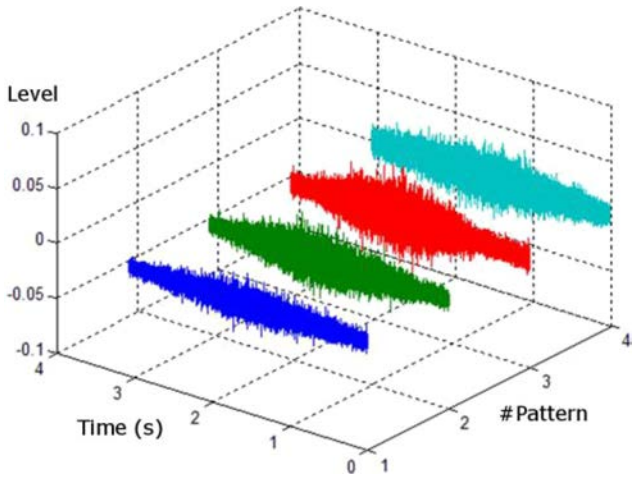

Fig. 4. Example of the result of the pass-by alignment process for four road vehicle sound samples.

5. Experimental evaluation

In this section, after describing the employed sound database, we explain the experimental setup so as to validate the performance of the proposed classification scheme.

5.1. Sound database

The considered sound database was built after carrying out a measurement campaign to record the noise sources in real environments. In concordance with (EU Directive, 2002) and (ISO 1996-2:2007), we considered the following noise source categories: light vehicles, heavy vehicles, motorcycles, aircrafts, trains and industrial noise. The recordings were performed using a Bruel & Kjaer 2250 sound level meter equipped with an integrated audio recording module, obtaining high quality recordings of 48 kHz, 16 bits and using a lossless coding system. In order to ensure the variability of the data, the noise sources were recorded in at least six

Table 1. Characteristics of the different types of noise sources recorded, where SUV stands for Sport Utility Vehicle.

Noise source	Characteristics	
Light vehicles	Car, SUV, van.	Urban streets and secondary roads
Heavy vehicles	Light trucks, heavy trucks.	
Motorcycle	Scooters (50 cc), motorbikes (> 125 cc)	
Aircrafts	Taking off and landing operations	
Trains	Commuter rail, regional rail, high speed rail, freight rail. Straight and curved railways.	
Industrial	Chimneys, machinery, cooling systems, etc.,	

different locations, each one presenting diverse conditions of background noise, distance to the noise source, etc. In the case of aircrafts, both landing and taking off operations were recorded at different distances to the flight path. Moreover, different kinds of trains were recorded, i.e., regional, high speed, freight trains, etc. The industrial noise samples were taken in the surroundings of several factories presenting different typologies: chimneys, machinery, refrigeration systems, etc. Finally, the road traffic vehicles were recorded in urban streets and secondary roads, trying to obtain clean vehicle pass-bys. Further details are provided in Table 1.

The resulting environmental noise database consists of 90 sound samples for each of the six noise source categories considered in this work (i.e. 540 noise samples). The duration of each sound sample was set to 4 seconds in order to consider the temporal evolution of the noise signals for their classification, as in (Chu *et al.*, 2009).

5.2. Experimental setup

Following the block diagram of Fig. 3, the signal features are pre-processed with Hamming windows of 30 ms and an overlap of 15 ms, as in (Valero, Alías, 2011b). Then, the thirteen signal features described in Section 3 are computed. MPEG-7 feature is referred to Audio Spectrum Envelope (ASE), since it is the MPEG-7 low-level descriptor that achieved the best results in previous works (Valero, Alías, 2010). In this work, as similarly done in (Ntalampiras *et al.*, 2008), we conduct a post-processing on this descriptor to improve its performance consisting of three steps: *i)* conversion into decibel scale, *ii)* energy normalization by RMS value, and *iii)* compaction of energy by the Discrete Cosine Transform. Cepstral-based features consider 13 coefficients, whereas Wavelet-based techniques consider 7 coefficients and are implemented using the "Daubechies" mother function, as in (Rabaoui *et al.*, 2004). Finally, notice that SBER is computed with four bands, as in (Sobreira *et al.*, 2008).

In order to select the optimal number of principal components to compress the merged feature vectors, experiments were run considering a sweep from 6 to 30 principal components for each pair of signal feature-machine learning technique. The number of components yielding the highest averaged recognition rates (between 8 and 16, depending on the case) was selected for the following experiments. All the machine learning techniques described in Sec. 3 (excluding the HMM) are asked to perform the classification of the sound patterns. To that effect, their configurations were adapted to the problem at hand, as explained in the following paragraphs.

The DT uses the Gini index as node split criterion since it works well with noisy data (Kim *et al.*, 2002) and has shown a good performance in similar works (Valero. Alías, 2011a). The KNN is implemented computing the Euclidean distance metric and considering $K = 3$ nearest neighbours, since it was the optimal value found in (Defréville *et al.*, 2006), being also used in (Sobreira *et al.*, 2008). Regarding the GMM, the related literature does not agree in a specific number of Gaussians: 40 were used in (Defréville *et al.*, 2006), 16 in (Ntalampiras et al., 2008) and 4-6 in (Chu *et al.*, 2009). Therefore we decided to empirically select this value by making a sweep between 5 and 40 Gaussians (with a step of 5 Gaussians). We selected 10 Gaussians since it was the value maximizing the classification accuracy. The implemented NN is a Multilayer Perceptron with only one hidden layer since, according to literature on NN, it is sufficient to approximate any given function (Cybenko, 1989). We keep the same NN configuration employed in our previous work (Valero, Alías, 2011b): 100 nodes in the hidden layer and 6 nodes (1 per class) in the output layer. Logarithmic sigmoid activation function is selected for all nodes, since the input data is previously normalized into [0,1]. NN weights are randomly initialised and Resilient Backpropagation learning algorithm is selected, given its good performance on pattern recognition problems (Riedmiller, Braun, 1993).

Finally, the classification accuracies are calculated as the percentage of correctly classified environmental noise samples, employing a 4-fold cross validation technique, where the 75% of the data is used for training and the 25% for testing, repeating the procedure 4 times with different training and testing sets (Chu *et al.,* 2009).

6. Results

In this section, we describe and discuss the results obtained from the conducted experiments. The first experiment evaluated the 52 possible combinations between the 13 signal features and the 4 machine learning techniques considered, in order to find the optimal one. The second experiment consisted in classifying noise samples of short duration, in order to validate the vehicle pass-by division into different phases. The third experiment evaluates the performance of the proposed classification scheme, besides comparing it to the baseline (i.e., the flat classification scheme of the first experiment). In the fourth experiment the comparison is preformed against the HMM (which already takes into account the signal time evolution with a flat classification scheme). Finally, listening tests are conducted to relate the achieved system accuracy to the human recognition ability.

6.1. Signal feature and machine learning method selection

First, we aim to select the most suitable combination of signal feature and machine learning technique for the problem at hand. For simplicity, experiments were run employing a flat classification scheme (without considering the vehicle pass-by phases). Results are shown in Fig. 5. It is worth noting that the DT yields the poorest performance regardless of the signal feature, as noted in (Valero *et al.*, 2011a). The rest of classifiers show a similar behaviour.

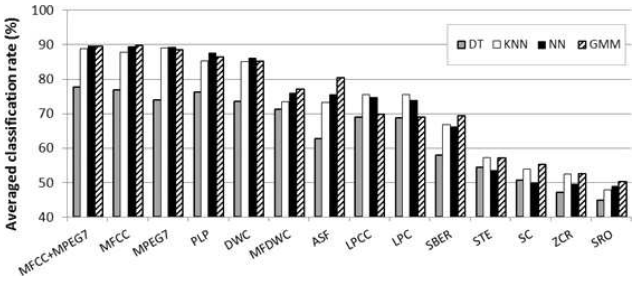

Fig. 5. Averaged classification accuracy obtained by the 13 signal features (plus the combination MFCC+MPEG7) when combined with the 4 machine learning techniques.

With regard to the signal features, there is a first group of descriptors that show very poor accuracy: ZCR, SRO, SC and STE. By means of these features the recognition system is able to hardly identify one out of two presented samples, proving a low ability to extract representative information from the noise signals. This can be explained by the inherent simplicity of these descriptors, since they all consist of a single coefficient per signal fragment. A second group of descriptors (SBER, SF, LPC, LPCC and MFDWC) show a notably better performance, presenting averaged classification rates close to 75%. In a third group, we found the signal features yielding the best performances: MPEG-7, MFCC, PLP and DWC. They all show good capabilities to extract relevant information from the noise signals at hand, attaining averaged recognition rates close to 90%. It should also be noted that a combination of the two best performing signal features (MFCC and MPEG7) was tested.

The highest absolute classification rates (with no significant differences) were attained by two different signal feature plus machine learning combinations: MFCC+MPEG7 plus NN; and MFCC plus GMM. Among them, the latter was selected for the rest of the experiments given its lower computation cost. The averaged recognition rate achieved by the system with that combination is 89.5%. The obtained confusion matrix confirms that, as in previous works (Defréville *et al.*, 2006) and (Valero, Alías, 2011b), the most common misclassifications occur between road vehicles (especially between heavy vehicles and both motorbikes and light vehicles) (see Table 2).

Table 2. Confusion matrix obtained from the flat classification scheme. In bold font, the most frequent confusions.

	Aircraft	Industry	Train	Light v.	Motorbike	Heavy v.
Aircraft	96.6	0.4	4.3	0	0	0
Industry	0	94.4	0.8	0	0	0
Train	2.5	2.1	92.3	0	0.3	2.6
Light v.	0	0	0.1	86.1	0.7	**8.6**
Motorbike	0.3	1.8	2.1	1.6	89.2	**10.2**
Heavy v.	0.6	1.3	0.4	**12.3**	**9.8**	78.6

6.2. Recognition of independent vehicle pass-by phases

The goal of the second experiment is to check the consistency of our pass-by division hypothesis. For that purpose, we divided each sound sample into three segments (approaching, passing and receding), considering every segment as an independent instance to be classified. In consequence, the size of the corpus is increased from 540 to 1080 samples. The recognition system was set to recognise not only the noise source, but also the phase of the pass-by in the case of road vehicles. MFCC and GMM were employed according to the results of the previous experiment.

As observed in the confusion matrix (see Table 3), the four most frequent confusions are between: *i)* motorcycle-approaching and heavy vehicle-approaching; *ii)* motorcycle-receding and heavy vehicle receding; *iii)* motorcycle passing and motorcycle-receding; and *iv)* light vehicle-passing and heavy vehicle-passing. Therefore, three out of the four most

Table 3. Confusion matrix (in %) obtained when considering the vehicle pass-by phases (A: Approach, P: Passing, R: Receding). Confusion rates below 1% are not shown. The grey boxes represent the phases belonging to the same road vehicle. In bold font, the most frequent confusions.

	Aircraft	Industry	Train	Light v.-A	Light v.-P	Light v.-R	Moto.-A	Moto.-P	Moto.-R	Heavy v.-A	Heavy v.-P	Heavy v.-R
Aircraft	99.3											
Industry		98.3						1.1				
Train			91.1	1.2	1.0		2.1	1.7	1.2	2.4	2.4	3.3
Light v.-A				88.6						7.8		
Light v.-P					85.0						6.6	
Light v.-R						86.8			1.3			7.0
Mot.-A			1.3				66.4	7.3	3.8	9.4	1.0	
Mot.-P			1.7		2.7		8.0	75.2	**11.1**		6.7	
Mot.-R			3.6			3.1	3.7	8.3	67.9		1.6	**15.2**
Heavy v-A				8.9			**17.1**	1.1		71.6	3.1	0.4
Heavy v.-P					**9.4**		0.2	4.4	1.2	5.9	71.7	2.6
Heavy v.-R			1.1			8.2			**11.8**	1.6	5.9	69.4

common misclassifications are produced between samples of different vehicles at the same pass-by phase, whereas only one is produced between different pass-by phases of the same vehicle. These results suggest that the pass-by phase is more discriminative than the type of vehicle itself, thus, validating the proposed approach that considers the vehicle pass-by phases independently for their recognition.

6.3. Performance of the proposed classification scheme

In this experiment, the performance of the proposed classification scheme is evaluated and compared to the flat baseline classification scheme (see Subsec. 6.1). The proposed scheme (see Fig. 3) uses a GMM to discriminate between road vehicles and not, and four independent GMM (one for the non-road vehicles and one for each of the three vehicle pass-by phases). The averaged classification accuracy obtained is 92.5%, which is 3% higher that the accuracy achieved by the flat classification scheme. In order to statistically validate the improvement achieved, the pairwise Student's t-Test is conducted. The result ($p = 2.5 \cdot 10^{-6}$) proves the improvement achieved.

If we compare the confusion matrix obtained with the proposed scheme (see Table 4) to the one obtained by the baseline (see Table 2), we can firstly observe the positive impact of the hierarchical structure on non-vehicle noise sources: the recognition of trains attains a relevant improvement of 5%. Secondly, the misclassifications produced between light and heavy vehicles are reduced to nearly the half (about 8%) thanks to considering the vehicle pass-by noise characteristics. Confusions between scooters and trucks are also significantly reduced, but in a minor degree (about 2.5%). Those confusion reductions result into an improvement of light and heavy vehicles classification rates, increasing the averaged accuracy in a 7% and a 4%in absolute terms, respectively.

Table 4. Confusion matrix (in %) obtained from the proposed classification scheme considering the vehicle pass-by phases.

	Aircraft	Industry	Train	Light v.	Motorbike	Heavy v.
Aircraft	97.3	0.2	0.1	0	0	0
Industry	0	95.8	0	0	0	0
Train	1.8	0.9	97.4	0	0.3	2.6
Light v.	0	0	0	93	0.8	7.2
Motorbike	0.2	1.9	2.4	1.3	89.1	7.8
Heay v.	0.7	1.2	0.1	5.7	9.8	82.4

6.4. Comparison to Hidden Markov Models

We wanted to compare the performance of the proposed classification scheme to the HMM classifier, which is one of the state of the art techniques applied for environmental noise recognition (Couvreur *et al.*, 1998; Rabaoui *et al.*, 2004) and (Ntalampiras *et al.*, 2008).

To this end, six HMM's were constructed (one per environmental noise source). The HMM yielding the highest log-likelihood (with respect to the unknown input signal) indicates the system output (i.e., recognised sound source). It should be noted that the HMM observations sequence corresponded to the sequence of feature vectors computed at every signal frame (i.e. no

feature merging was applied since HMM already model the time evolution of input signals). The outputs from each HMM were modelled with a mixture of 10 Gaussians (as set with the GMM in the previous experiments) initialized with the K-means algorithm. The HMM parameters (transmission and emission probabilities) were estimated by using the Baum-Welch algorithm, as in (Ntalampiras *et al.*, 2008). Two HMM topologies (i.e. fully-connected and left-to-right) with several number of hidden states were analysed. As shown in Fig. 6, the left-to-right typology yields better performance than the fully-connected structure, which suits to the characteristics of the environmental noise events (Rabaoui *et al.*, 2004). Specifically, the left-to-right HMM with 3 states yields the highest averaged classification accuracy. This configuration is in concordance with the proposed scheme, which characterizes a vehicle pass-by in 3 phases (equivalent to the 3 states) without the possibility of backward transitions (left-to-right).

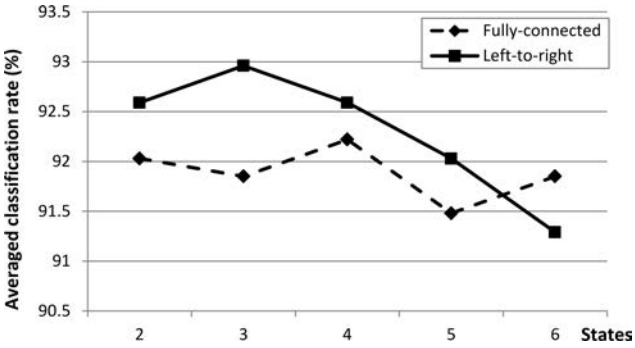

Fig. 6. Averaged recognition rates yielded by two topologies of HMM for different number of hidden states.

With this optimal configuration, the HMM with a flat recognition scheme obtains an averaged classification rate of 92.9%, which is 0.4% higher than the proposed scheme. However, this improvement is not statistically significant according to the pairwise Student's t-Test ($p = 0.41$).

6.5. Listening tests

A set of listening tests were conducted to compare the performance of the system with the human ability to recognise these kinds of environmental noise sources. We employed the multimedia testing platform called TRUE (Planet *et al.*, 2008). The test employed 120 sound samples extracted from the corpus used in the previous experiments. A total of 30 subjects completed the test, from which 15 were experts from music, speech or other audio-related fields and only one was expert on environmental noise recognition tasks. The subjects completed the tests from their home and using their own headphones. They were asked to perform the tests in low background noise conditions and maintaining the same reproducing volume during the whole test.

Two different listening tests were conducted. In the first one, 60 sound recordings of 4 seconds long were presented to the subject in a forced-choice test (*unknown* answer was not available). The sound files, 10 per noise source, were randomly selected from the corpus. The averaged recognition rate attained by human listeners is 80.3%, which is significantly lower than the 92.5% achieved by the trained system. As it may be concluded from Fig. 7a, the responses of the evaluators show a large variability: a well-trained expert listener can achieve a 100% of recognition accuracy, but less trained or non-expert listeners may only correctly recognise 2 out of 3 noise sources. On the contrary, the standard deviation of the performance attained by the automatic recognition system is notably lower.

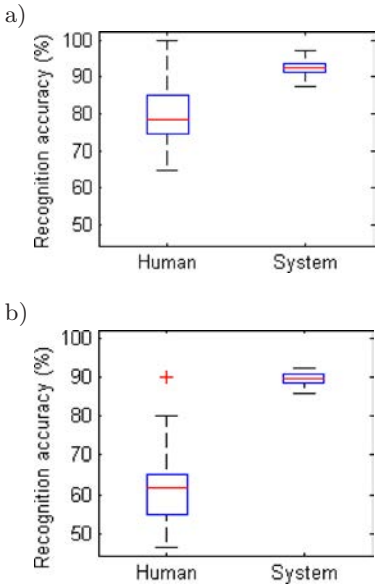

Fig. 7. Boxplot of the recognition results attained by both the human and proposed classification system for the first (a) and second (b) listening test.

In the second listening test, the sound samples were shorter (exactly 1.33 seconds, as considered in the experiment from Subsec. 6.2). 60 sound samples composed the test, ensuring an equal distribution between noise sources and also between pass-by phases. The test results show a dramatic decrease in the human recognition performance with respect to the previous experiment, attaining 62.2% of recognition accuracy on average (see Fig. 7b). After these results, it is clear that the time envelope of the signal is of paramount importance for the recognition of this kind of noise events in human beings, as noted in (Gygi, 2001). On the contrary, the proposed recognition system does not need that much information to perform the identification: the accuracy in that case remains at very high recognition rates (close to 90% in average).

7. Discussion

After analysing the results obtained by considering the different combinations of signal features and machine learning techniques, we can conclude that there are mainly two sound signal features (MPEG-7 and MFCC) and three machine learning techniques (KNN, NN and GMM) that may yield to high classification accuracies. However, if we aim to go a step forward in terms of performance, a structural change in the classification scheme should be introduced. In that sense, the proposed classification scheme takes advantage of the spectro-temporal characteristics of the vehicles pass-bys in combination with a hierarchical classification structure, yielding a significant increase (+3%) from an averaged classification rate which is already quite high (89.5%). Furthermore, the results extracted from experiments employing short duration samples (containing only one phase of the vehicle pass-by) validate the consistency of the initial hypothesis of considering the vehicle pass-by phases as independent classes (since further confusions are found between samples from different vehicles at the same pass-by phase than between different pass-by phases of the same vehicle). This ability will be particularly valuable when exposing the system to situations where full pass-bys cannot be clearly identified (e.g., streets or roads with greater traffic density). In that case, it is foreseen that the recognition system will be able to provide a robust recognition decision (i.e., noise source) quite fast and only based on a short fragment of the noise signal pass-by (which could correspond to the approaching, passing or receding phase).

Moreover, in order to complete the experimental analysis, the proposed classification scheme was compared to HMM. As HMM inherently takes into account the temporal evolution of the sound events, it yielded an equivalent performance. However, the proposed classification scheme provides two advantages with respect to HMM for the problem at hand. First, it has a notably lower computational cost, since it only needs three observations (corresponding to the 3 pass-by phases) to classify the noise source, whereas the HMM uses one sequence per signal frame (in this work, 266 sequences[1]). And second, the proposed scheme contains observable states, i.e. they are not hidden (i.e., *approaching, passing-by* and *receding* phases), thus having further information about the vehicle position at any instant of time.

Finally, the listening tests provide a reference with regard to the human ability to recognise noise events, showing that an average listener is not able to perform as well as the proposed system if he/she is not specifically and exhaustively trained for that purpose, as the system is. In consequence, the classification system outperforms the average human listener in a 12%. The second part of the listening test (employing short duration samples), stresses the robustness of the proposed recognition system: while the human ability decreases on 18% with respect to using the long 4 s samples, the accuracy of the system only drops in 4%. This comparison agrees with the results reported in (Couvreur *et al.*, 1997), where also human listeners attained a lower recognition accuracy that the trained recognition system.

[1]The number of sequences is obtained as the noise sample length (in this work, 4s) divided by the signal frame step (in this work, 15 ms).

8. Conclusions

This paper has addressed the recognition of the environmental noise sources typically encountered in urban areas, which may affect the citizen's quality of daily life. After determining the best signal feature and machine learning technique combination for the corpus at hand, we have proposed a classification scheme that takes into account the noise signal characteristics of road vehicles pass-bys. Experiments conducted on a corpus of environmental noise samples recorded in real environments show a significant improvement in the classification accuracy of the proposed classification scheme when compared to a traditional flat scheme, in particular, by decreasing the confusions between light and heavy vehicles. An 8% reduction is obtained, which is attributed to the differences in the pass-by phases from both noise sources.

When compared to the state of the art HMM, the proposed classification scheme achieves an equivalent accuracy performance but with a significant lower computational cost, besides providing directly observable information about the current state (i.e., phase) of the vehicle pass-by. The performed listening tests have highlighted the excellent recognition accuracy achieved by the proposed system: an average non-trained human listener attains noise source recognition accuracy significantly lower than the proposed system (about 10% in average for samples containing the whole vehicle pass-by and about 25% for samples containing only one particular phase of each vehicle pass-by).

Future work will be mainly focused on facing the detection of noise events within continuous signals. Also, the proposed classification scheme will be extended so as to study overlapped road vehicle pass-bys by considering three subclasses for each noise source (referred to each of the pass-by phases), which will allow a more flexible and fast-response system.

Acknowledgment

Part of this work was presented at Tecniacústica-2011 (26–28 Oct, Cáceres, Spain).

References

1. Amman S.A., Das M. (2001), *An efficient technique for modeling and synthesis of automotive engine sounds*, IEEE Trans. Ind. Electron., **48**, 225–234.
2. Babisch W. (2006), *Transportation noise and cardiovascular risk: Updated review and synthesis of epidemiological studies*, Noise&Health, **8**, 30, 1–29.
3. Bishop C.M. (2003), *Neural Networks for Pattern Recognition*, Oxford Univ. Press, New York.
4. Breiman L., Friedman J., Olshen R., Stone C. (1993), *Classification and Regression Trees*, Chapman&Hall, Boca Raton.
5. Cevher V., Chellappa R., McClellan J.H. (2009), *Vehicle speed estimation using acoustic wave patterns*, IEE Transactions Signal Processing, **57**, 1, 30–47.
6. Chu S., Narayanan S., Jay Kuo C.-C. (2009), *Environmental sound recognition with time-frequency audio features*, IEEE Transactions Audio, Speech and Language Processing, **17**, 2, 1142–1158.
7. Couvreur C., Fontaine V., Gaunard P., Mubikangiey C.G. (1998), *Automatic classification of environmental noise events by Hidden Markov Models*, Applied Acoustics, **54**, 3, 187–206.
8. Cybenko G. (1989), *Approximations by superpositions of a sigmoidal function*, Mathematics of Control, Signals and Systems, **2**, 303–314.
9. Defréville B, Roy P., Rosin C., Pachet F. (2006), *Automatic recognition of urban sound sources*, Proceedings of the 120th AES Convention, Paris.
10. Eronen A., Peltonen V., Tuomi J., Klapuri A., Fagerlund S., Sorsa T., Lorho G., Huopaniemi J. (2006), *Audio-based context recognition*, IEEE Transactions Audio, Speech, Lang. Processing., **14**, 1, 321–329.
11. EU Commission (1996), *The Green Paper on Future Noise Policy*, (COM(96) 540).
12. EU Directive (2002), *Directive 2002/49/EC of the European parliament and the Council of 25 June 2002 relating to the assessment and management of environmental noise*, Official Journal of the European Communities, L 189/12, July 2002.
13. Gygi B. (2001), *Factors in the identification of environmental sounds*, Ph.D. Thesis, Indiana University.
14. ISO (2001), *ISO/IEC FDIS 15938 4:2001, Information Technology Multimedia Content Description Interface – Part 4: Audio.*
15. ISO (2007), *ISO 1996-2:2007 Acoustics – Description, measurement and assessment of environmental noise – Part 2: Determination of environmental noise levels.*
16. Jones M.T. (2008), *Artificial Intelligence – A Systems Approach*, Infinity Science Press, Higham.
17. Kim Y., Jeong S., Kim D. (2002), *A GMM-based Target Classification Scheme for a Node in Wireless Sensor Network*, IEICE Trans. Fundamentals/ Commun./Electron/Inf&Syst., **E85**, 1.
18. Kim H., Moreau N., Sikora T. (2005), *MPEG-7 Audio and beyond. Audio content indexing and retrieval*, [Ed.] John Willey & Sons Ltd., Chichester.
19. Ntalampiras S., Potamitis I., Fakotakis N. (2008), *Automatic Recognition of Urban Environmental Sound Events*, Proceedings International Association for Pattern Recognition Workshop on Cognitive Information Processing, Santorini.
20. Peeters B., Blokland G. (2007), *The Noise Emmission Model for European Road Traffic*, Deliverable 11 IMAGINE project, EU 6th FP.
21. Planet S., Iriondo I., Martínez E., Montero J. (2008), *True: an online testing platform for multimedia evaluation*, Proceedings 2nd International Workshop on EMOTION: Corpora for Research on Emotion and Affect at LREC08, Marrakech.
22. Rabaoui A., Lachiri Z., Ellouze N. (2004), *Automatic Environmental Noise Recognition*, Proceedings IEEE International Conference on Industrial Technology, Hammamet.
23. Rabiner L., Juang B-H. (1993), *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ.
24. Rabiner L. (1989), *A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceedings of the IEEE, **77**, 2, 257–286.
25. Rasche F. (2004), *Arousal and aircraft noise – Environmental disorders of sleep and health in terms of leep medicine*, Noise & Health, **6**, 22, 15–26.
26. Riedmiller M., Braun H. (1993), *A direct adaptive method for faster backpropagation learning-RPROP algorithm*, Proceedings IEEE International Conference on Neural Networks, Nagoya.
27. Sandberg U., Ejsmont A.J. (2002), *Tyre/Road Noise Reference Book*, Kisa, Sweden.
28. Sobreira Seoane M., Rodriguez Molares A., Alba Castro J.L. (2008), *Automatic classification of traffic noise*, Proceedings Acoustics'08, Paris.
29. Umapathy K., Krishnan S., Jimaa S. (2005), *Multigroup classification of audio signals using time-frequency parameters*, IEEE Trans. Multimedia, **7**, 2, 308–315.
30. Valero X., Alías F. (2010), *Applicability of MPEG-7 low level descriptors to environmental sound source recognition*, Proceedings 1st Euroregio Conference, Ljubjana.
31. Valero X., Alías F. (2011a), *Comparison of machine learning technique for the automatic recognition of soundscapes*, Proceedings Forum Acusticum'2011, Aalborg.
32. Valero X., Alías F. (2011b), *Automatic monitoring of environmental noise sources*, Proceedings Tecniacustica'2011, Cáceres.