Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Universal background modeling for acoustic surveillance of urban traffic

Stavros Ntalampiras

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, 20133, Italy

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Acoustic signal processing
Generalized sound recognition
Universal background modeling
Hidden Markov model
Wavelet packets
Intelligent transportation systems

ABSTRACT

Traffic congestion in modern cities is an increasing problem having significant consequences in our daily lives. This work proposes a non-intrusive, passive monitoring framework based on the acoustic modality which can be used either autonomously or as a part of a multimodal system and provide valuable information to an intelligent transportation system. We consider a large number of audio classes which are typically encountered in urban areas. We introduce a combination of a powerful audio representation mechanism based on time, frequency and wavelet domain features with universal background modeling which leads to higher recognition accuracies and detection rates (in terms of false alarm and miss probability rates) with respect to commonly employed methodologies. The basic advantage of a class-specific model derived using the universal background modeling logic is its tolerance to data which belong to other sound classes. Another important feature of the proposed system is its ability to detect crash incidents, which apart from their catastrophic impact on human life and property, have negative consequences on the traffic flow. Our experiments are based on the concurrent usage of professional sound effect collections which include audio recordings of high quality. We thoroughly examine the performance of the proposed system on isolated sound events as well as continuous audio streams using confusion matrices and detection error trade-off curves.

1. Introduction

Modern societies in general and especially major cities are currently facing the constantly increasing problem of traffic congestion. A characteristic statistic is that the number of vehicles which traveled within the U.S. borders between 1980 and 1998 increased 72% while at the same time interval the number of lane lines increased only 1% [33]. Since the addition of more lanes is becoming less and less feasible, contemporaneous approaches are concentrated on using the existing infrastructure more efficiently based on reliable traffic information. In order to limit the specific problem several Intelligent Transportation System (ITS) programs have been initiated, e.g. the Federal ITS program by the U.S. Government in 1991. The overall aim is the development of technologies which maximize the traffic capacity of a certain area and concurrently minimize the delay of the transportation. However, the current technological solutions cannot meet the traffic demands with respect to most major urban areas.

In general, ITS systems are employed for: a) information provision to travelers using the variable message sign (VMS) technology, b) free-way and arterial management, c) crisis situation management, d) parking management and d) increase traffic safety. It is evident that the particular systems heavily rely on the traffic conditions of the network. Typically an ITS employs electronic, computer and communication technologies into vehicles and roadways for increasing traffic safety and reducing congestion under the prism of improving the quality of life of the general public. In this context, information about the traffic conditions is extremely useful, thus there is an immediate need for automated monitoring of urban traffic in order to facilitate ITS methodologies. For improved ITS performance, monitoring should operate in real-time, be diverse (different kind of information should be available) and provide reliable scene analysis under a wide range of environmental conditions. We suggest that traffic surveillance may be based on the acoustic emissions of each vehicle or urban environmental sound event, a process which is non-invasive and involves signal processing and pattern recognition algorithms. Even though the current traffic surveillance systems are mostly based on optical information (e.g. Smart-Cam [3,20,11,10]), the acoustic modality could be used in parallel for improved performance and better quality of service, following the logic of [26].

The motivation behind an urban traffic monitoring system based solely on the acoustic modality is that it may assist the measuring of various traffic parameters such as flow and density of

E-mail addresses: sntalampiras@upatras.gr, dalaouzos@gmail.com.

http://dx.doi.org/10.1016/j.dsp.2014.05.003
1051-2004/© 2014 Elsevier Inc. All rights reserved.

vehicles inside the area of interest. Subsequently these parameters can be used by experts through a decision support interface for reducing congestion and/or pollution by applying a more appropriate traffic management plan. Alternative routes may be suggested to the drivers and potentially save valuable time and energy. Overall, such monitoring systems allow for a more efficient usage of existing infrastructures, which nowadays is the main target as road building is no longer seen as an acceptable solution [16].

This type of monitoring may quicken traffic incident response times and expedite the recovery process. It could also be proven useful to manage the urban development by taking into account the sonic information and constructing sound maps for categorizing and indexing sources of urban noise pollution [31,2]. Other applications include informing travelers about the possibility of heavy traffic or more efficient routing (direct traffic to particular routes or areas or give priority to specific categories of vehicles), pavement maintenance, adaptive road signal management based on real-time traffic flow, etc.

After providing a thorough review of the related work, we experiment on the classification of nine categories (*car*, *motorcycle*, *aircraft*, *crowd*, *thunder*, *wind*, *train*, *horn* and *crash*) based on the experience gained from [29]. The novel aspects of the proposed work are the following:

1. The utilization of a multidomain feature set (time, frequency and wavelet) in the context of urban audio signal processing.
2. The classification stage relying on generative models which are adapted versions of a universal one. The universal background model (UBM) is an effective framework and has received little attention from the generalized audio recognition community.
3. The proposition of a reference dataset which is appropriate for the specific task and will permit the reliable comparison of approaches with similar goals.
4. Last but not least, the incident detection component (to the best of our knowledge a work which addresses the particular problem has not been previously addressed). The particular component caries significant importance as it can notify the authorized manager of a potentially hazardous situation and may assist limiting its consequences including the activation of an alternative traffic management plan.

The rest of this article is organized as follows: Section 2 provides an overview of the related literature while the emphasis is being placed on approaches which include audio signals belonging to vehicles often encountered in urban areas. Section 3 analyses the modules which comprise the proposed surveillance framework with special attention to the universal background modeling. The next two sections (5 and 6) examine the detection capabilities of the proposed approach in a thorough and concise way. Finally Section 7 offers our conclusions as well as ideas for future works.

2. Related literature

Several studies have been conducted which fall into the general area of processing of urban audio signals. However there are no many studies which address the problem through the traffic management point of view nor they consider a large number of audio classes as the present work does. An interesting approach is described in [22] where the continuous usage of features coming from the time, spectral and cepstral domains is applied for the distinction of cars, vans and trucks. Their classification system is based on the Support Vector Machine algorithm. Another approach is explained in [40] where the "eigenfaces method" (borrowed from the field of face recognition) is employed. The main characteristic is that the frequency spectrum of about 200 ms is treated as a vector in a high-dimensional frequency feature space. When a new vector is processed, its spectrum is compared to the already processed spectrums and the difference vector is projected onto the principal component axes to compute the residual. Their experiments are focused on discriminating cars from other vehicles. A Classification and Regression Tree (CART) classifier is proposed in [1] which is based on the minimal distance. The input to the CART is an acoustic signature representative of the distribution of the energies among blocks which consist of its wavelet packet coefficients. The categories which are considered are cars, trucks and vans. Another approach which tries to create sound maps of the monitored area based on passive detection of sounds emitted by road vehicles is presented in [7]. The particular method does not perform any kind of vehicle identification (and thus is only indirectly connected to the present article) but rather tries to approximate parameters reflecting the road conditions, e.g. vehicle count, speed, etc. In the end, the authors suggest that further research is required to compute these parameters in a secure and reliable manner from the sound field maps. An approach which is based on Gaussian mixture models (GMM) is detailed in [28]. The focus is placed upon the classification of two classes, light vs. heavy vehicles where light vehicles include cars, SUVs, minivans and light trucks while the heavy ones include heavy diesel trucks and buses. Their feature set is formed by the output energy levels of a generalized parametric non-linear filterbank. The GMM method is compared to a linearly weighted discriminator and it is concluded that they may be used in a collaborative manner.

A narrower problem which falls into the category of urban audio signal processing is addressed in [15]. The authors used the Time Encoded Signal Processing and Recognition (TESPAR) method combined with the archetypes technique. A variety of Butterworth low pass filters was explored while their recordings included six different models of cars. In [13], the authors explain the usage of time-varying autoregressive (TVAR) modeling to analyze acoustic signatures of six different classes of moving vehicles. They employ an artificial neural network for the classification fed with the TVAR parameters expanded by a low-order discrete cosine transform. Another interesting work presented in [17,18] uses different classifiers (MLPs) trained on individual noise types in the context of computational auditory scene analysis.

It is worth mentioning that there is also a substantial amount of work done in the area of classification of acoustic emissions coming from *military* vehicles (e.g. [39,27]). However the associated signals exhibit special properties and therefore they are treated differently in terms of features and classifiers. Another line of thought is followed in [22] where the problem of separating large trucks, small trucks and cars is dealt by the concurrent usage of audio and visual information. The merits of fusing data on feature level are explained since they allow to decrease the number of learning samples in order to obtain the same classification accuracy with mono-modal data. In this work we are based solely on the acoustic modality which in many cases may provide information that is difficult or even impossible to capture by any other means while, in general, algorithms of lower computational complexity are involved. Furthermore the present article considers signals which may help the vehicle identification task, unlike [37] where the focus is on cues present in the cumulative acoustic signal acquired from a roadside installed single microphone.

3. System blocks

The specific section describes the two main modules of the proposed surveillance system. It is divided into two parts following the fundamental logic behind generalized sound recognition, which states that each sound source distributes its energy across different frequencies in a unique way. Initially we try to capture this way by extracting acoustic features and then model them using statistical

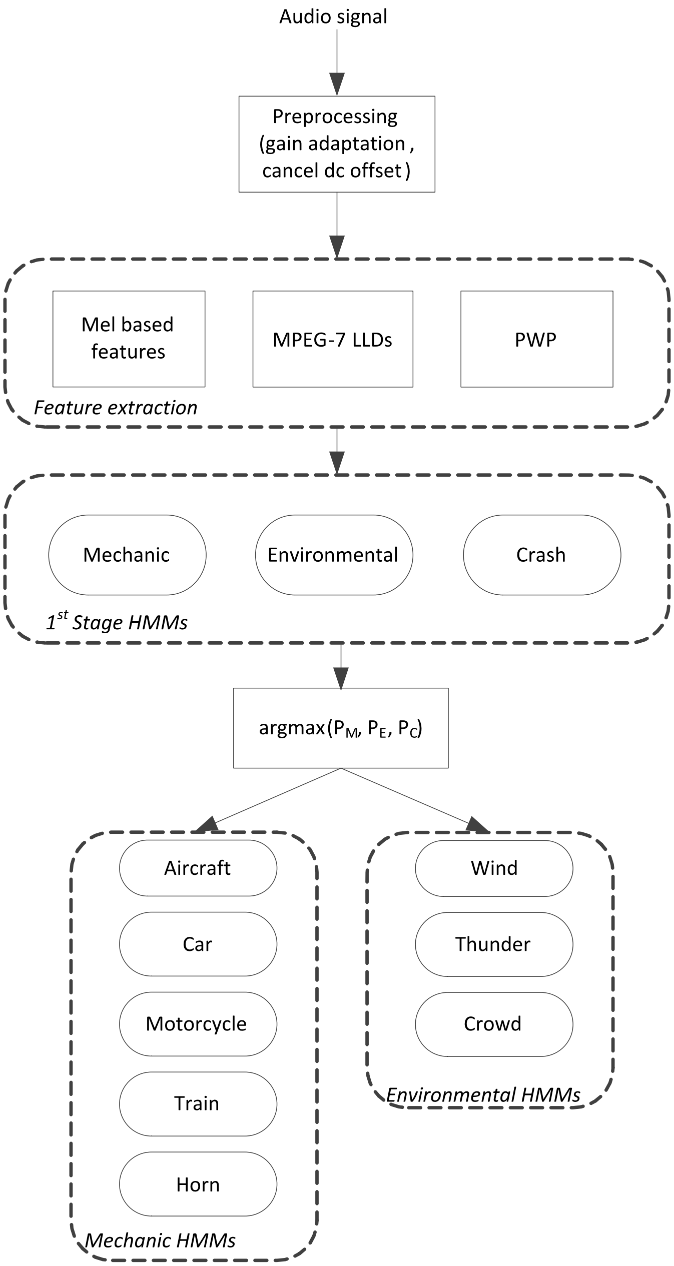

Fig. 1. The block diagram of the two-stage topology for acoustic surveillance of urban environments.

pattern recognition algorithms. The first part of this section is devoted to the analysis of the feature extraction processes while the second part is concentrated on the pattern recognition approach.

The overall block diagram is depicted in Fig. 1. The signal is initially preprocessed including gain normalization and mean removal. This phase ensures the exploitation of the entire audio spectrum for facilitating further processing and avoiding any information loss. Subsequently three feature extraction algorithms are applied: *a*) Mel-frequency cepstral coefficients, *b*) MPEG-7 low level descriptors, and *c*) perceptual wavelet packets. They represent the current trend on generalized audio recognition quite well, a fact which motivated their usage. Moreover, they include descriptors belonging to various domains (time, frequency and wavelet) towards capturing diverse aspects of the structure of the audio signal. The following step includes probabilistic model-based classification into *mechanic*, *environmental* and *crash* sound events. We apply HMM based modeling in the following two versions: *a*) class-specific modeling: each model is created using data coming from one class alone and, *b*) universal background modeling: information from the whole dataset is used for constructing the class-specific models. Finally, the path producing the maximum log-likelihood is followed and the novel sound is labeled with a specific class.

3.1. Audio features

This paragraph provides only a brief description of the groups of sound parameters which were employed. For a detailed analysis the interested reader is directed to a previous work of ours [30]. The groups of sound parameters are the following:

- *Mel-Frequency Cepstral Coefficients (MFCC)*: The particular group originates from the speech/speaker recognition community. 23 Mel filter bank log-energies were utilized. After the windowization of the signal the short time Fourier transform (STFT) is computed for every frame while its outcome is filtered by a triangular Mel scale filterbank. The logarithm is then obtained to space the data and finally the discrete cosine transform (DCT) is applied, a process which decorrelates the data while in practice it is as efficient as the data-driven projection techniques at a much lower computational cost.
- *MPEG-7 Low Level Descriptors (LLDs)*: The MPEG-7 audio protocol aim at the establishment of a generic methodology for processing generalized audio signals [5]. To this end it offers not only standardized LLDs but also Descriptions Schemes where the input audio information is characterized at a higher level.
- *Perceptual Wavelet Packets (PWP)*: The previously mentioned acoustic parameters are derived either from time or frequency domain. The specific set takes advantage of the unique properties of the wavelet domain. The main advantage of the wavelet transform is that it can process time series, which include non-stationary power at many different frequencies [36]. While sinusoids are smooth and predictable, wavelets tend to be irregular and asymmetric.

The aforementioned groups were initially used separately in order to assess their performance on the task of urban traffic surveillance. A series of experiments was conducted in order to select the combination of parameters which provided the highest recognition rates. Even though some of these features may provide redundant information, one should keep in mind that this may not be a totally undesirable property since in many cases the audio signals may be distorted or even some parts of the spectrum may be totally absent. This could be a result of heavy noise or other causes which are hard to predict or know before having the system operating in the desired environment. It should be mentioned that during the last phase of the computation of the MPEG-7 Audio Standard LLDs as well as PWP features, the DCT was used in order to reduce the dimensionality of the feature vector.

3.2. Audio pattern modeling

The distribution of the audio parameters is approximated using the hidden Markov model (HMM) approach, which are probabilistic finite state machines providing encouraging performance in audio recognition tasks [21]. Each state of the HMM is modeled by a Gaussian mixture model with diagonal covariance matrix, which in practice is equally effective to the fully-covariance one [25]. In brief HMMs are sequential statistical models where the underlying states are hidden but have a stationary distribution leading to their remarkable modeling capacity in case of sequential data such as speech, image, video, financial data, etc. Initially, the HMM technique divides the feature sequence into states. Subsequently each state is learnt using a GMM while the associations between the GMMs are represented by a transition probability matrix, which includes the probability for all possible transitions across the states. Left–right HMMs, i.e. models which permit only left to right transitions are typically employed for generalized sound recognition tasks due to the sequential nature of most sounds. This work applies and compares the following two HMM logics:

- *Class specific HMMs*: during this phase we create one HMM to represent each sound class (using data associated with the specific class alone) and we follow the left–right topology.
- *Universal HMM*: during this phase one HMM is created based on the entire training dataset while adapted versions of it are used to represent each sound class. In this case we use fully-connected (or ergodic) HMMs where every possible transition is permitted by the model which comprises a more appropriate choice given the variability of the entire dataset.

The following section explains in more detail both the methodology and the motivation behind the particular modeling approach.

3.3. Universal background modeling

The basic idea behind the creation of a universal HMM is to provide robustness against mismatched data which otherwise would have been unseen. The UBM is a class-independent model trained on a great gamut of data or by pooling sub-population models which are trained on selected parts of the entire dataset [34]. A problem which is addressed by the UBM logic is the low quantity of data with respect to one or more categories. Thus the models of those classes can be derived by adapting the parameters of the universal model using fewer training data. Finally the problem of the recording conditions can be solved by the usage of a large model which is independent of the recording conditions given that it is pre-trained using a plethora of data belonging to each condition (reverberation, noise, etc.). For example, if one creates the UBMs for each recording condition a-priori, the current condition can be selected according to the *Maximum-a-Posteriori* (MAP) [14] principle for practical applications. The motivation behind adaptation is to derive the class specific model by updating the well-trained parameters of the UBM via adaptation. This provides a tighter coupling between the model of each category and the UBM.

In this work we use solely mixtures of Gaussian with diagonal covariance matrices mainly because a) an Mth order full covariance GMM can be replaced by a larger order of diagonal GMM with equal modeling capabilities [34], b) diagonal GMMs require significantly less computational resources and c) based on early experiments, diagonal GMMs provided higher recognition rates compared to full-covariance GMMs.

It is of critical importance to train the UBM on a dataset which is balanced across the categories we wish to model. In the opposite case, the constructed model will be biased towards the dominant subpopulation. In this work we employ the MAP method for adapting the UBM which compared to the Maximum Likelihood Linear Regression results at more accurate modeling when a relatively large quantity of adaptation data is available [41]. *MAP* achieves two goals directly related to the recognition capabilities of the system: a) model components which are seen inside the adaptation data are emphasized while b) the remaining components are deemphasized. As a result when the test data include components which in the class-specific model would have been unseen, the system is not confused. Furthermore we added a relevance factor which indicates how much new data have to be observed in a mixture before the current parameters begin to update. Thus when limited data correspond to a specific mixture its parameters are not replaced.

There are different directions that can be followed in order to use the data for the construction of the UBM. One is to simply pool all the (balanced) data and train the model. Another direction is to train different UBMs representative of a specific subpopulation of the corpus and subsequently merge the models into one. The basic advantage of this approach is that one may use unbalanced data and still come up with a reliable model since one is able to control the composition of the final UBM. Some other directions have been followed in the literature (see for example [35] and [19]). In this work we collected data which are uniformly distributed among all the involved sound classes for building the universal model. Subsequently we employed the rest of the training sequences for adapting it and coming up with class-specific models.

For completeness we will briefly describe the adaptation process which was followed. *MAP* adaptation requires prior knowledge regarding the distribution of the model parameters (usually called informative prior), thus exploiting adaptation data of restricted quantities in an effective way. In our case the informative priors are the parameters of the universal class-independent model. Young et al. [41] use conjugate priors for mathematical tractability and derive a relatively simple adaptation scheme. Each mean μ of the already constructed model is updated based on the next formula:

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\bar{\mu}_{jm} + \frac{\tau}{N_{jm}+\tau}\mu_{jm}, \tag{1}$$

where j denotes the state of the HMM, m the Gaussian mixture, τ a weighting factor, $\bar{\mu}$ the mean of the sound parameters of the adaptation data and $\hat{\mu}$ the updated mean. N comprises the occupation likelihood with respect to the adaptation data and is given by:

$$N_{jm} = \sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t), \tag{2}$$

where R is a specific training sequence (part of a known audio signal), T_r is the observation at time t and L the likelihood produced by the m-th mixture of the j-th state with respect to sequence r at time t. The mean $\bar{\mu}$ of the incoming adaptation data is computed using the following formula:

$$\bar{\mu} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)o_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)}, \tag{3}$$

where o_t^r denotes the incoming observation of data which belong to the adaptation sequence r at time t.

As a result when N_{jm} of a specific Gaussian component has a high value, its MAP estimate will change dramatically. On the contrary, it will be almost equal to the universal one. Lastly, it should be noted that as long as the relevance factor is reached, every mean of every Gaussian function is updated based on the prior mean, the weighting and the adaptation feature vector.

4. System development

This section provides a thorough description of the procedure that was followed in order to design the final form of the traffic surveillance framework. The database included the following nine audio classes: *aircraft*, *car*, *motorcycle*, *crowd*, *thunder*, *wind*, *train*, *horn*, and *crashes*. We thoroughly searched several professional sound effect collections (e.g. BBC Sound Effects Library, Sound Ideas Series 6000, Sony Sound Effects Library, etc. [12]) to locate all the related events. This type of datasets contains enormous quantities of high quality audio events and are ready to be used by the scientific community. Our main interest was for the sample to be free of any type of background noise. An identical dataset to any of the ones reported in the literature could not be established, mainly because this article considers a wider range of audio categories. There exists an absence of a reference database even for the common classes, thus our work is not directly comparable to the previous ones. However we contrasted the proposed universal modeling of multidomain parameters approach to the one commonly employed by the audio recognition community. That includes the modeling of the feature set by multiple HMMs, each one representing one audio class [8,23,6].

The final corpus was divided into 70% for training and 30% for testing each experimental scheme while the division was kept the same during all phases. In addition the sound files were downsampled to 16 kHz with 16-bit quantization while they were preprocessed in order to cancel any possible DC-offset. Table 1 tabulates descriptive statistics of the final dataset. It is important to note that there is no any kind of overlap between the train and the test set at any processing stage.

The first stage of the system development phase examined the topology of the classification scheme. Two topologies were tested while we used the MFCC set of parameters. The first one was a single stage topology where one statistical model was built for each class. Here, classification is performed simply by evaluating the log-likelihoods produced by the model of each class and selecting the model generating the maximum one (this is similar to the so-called direct approach described in [21]). The second one was based on our previous experience as well as on the a-priori knowledge we have about the audio classes. Two stages were designed in order to limit the problem space based on the way which humans categorize subconsciously their surrounding environment based solely on the acoustic information (see also Fig. 1). The specific approach tackles one of the major problems that sound recognition technology has to face which is the performance decrement as the number of categories increases [32].

The confusion matrices related to the first system development phase are shown in Tables 2–5. It should be mentioned that during every phase of our experimentations, the parameters of the hidden Markov models were drawn from the following sets: number of states: $\{3, 4, 5, 6, 7\}$ and number of Gaussian components: $\{2, 4, 8, 16, 32, 64, 128, 256, 512\}$. The final selection was made based on the highest recognition rate criterion while the recognition rates were averaged across all classes so that the final average value is not biased favoring the class which dominates the corpus.

The average recognition rates for the single and two-stage systems are 61.4% and 66.2% respectively, which comprises a significant increment. Thus we decided to employ the latter approach for our further experiments.

Table 1
Statistics of the dataset composed of professional sound effect collections.

Audio class	# of samples	Total duration (s)
Aircraft	110	4481.8
Motorcycle	79	1222.5
Car	81	2200.9
Crowd	60	3312.2
Thunder	61	977.1
Wind	66	3232.6
Train	82	2455.1
Horn	194	817.4
Crash	149	1975.2
Total	**882**	**20 674.8**

Table 2
The confusion matrix with respect to single stage topology based on the MFCC set.

Presented	Predicted (%)								
	Aircraft	*Motorcycle*	*Car*	*Crowd*	*Thunder*	*Wind*	*Train*	*Horn*	*Crash*
Aircraft	**55.8**	5.1	15.6	5.3	12.8	0	2.5	2.9	0
Motorcycle	8.3	**50**	0	0	8.5	0	16.6	0	16.6
Car	10.3	11	**48.2**	10	3.4	0	13.7	3.4	0
Crowd	0	0	0	**86.4**	0	0	9.1	4.5	0
Thunder	9.8	6	7	3	**43.3**	0	16.5	11.4	3
Wind	3	4	2	4	16	**52**	8	3	8
Train	9.6	4.8	0	13.9	2.4	4.9	**64.4**	0	0
Horn	0	2.4	0	5.1	3.2	2.1	7.4	**75.6**	4.2
Crash	1.8	0	5.7	0	11.9	0	3.6	0	**77**

Table 3
The confusion matrix with respect to the first phase of the two-stage topology based on the MFCC set.

Presented	Predicted (%)		
	Mechanic	*Environmental*	*Crash*
Mechanic	**78**	10.6	11.4
Environmental	5.4	**87.2**	7.4
Crash	2	11.2	**86.8**

Table 4
The confusion matrix with respect to the mechanic group of the two-stage topology based on the MFCC set.

Presented	Predicted (%)				
	Aircraft	*Motorcycle*	*Car*	*Train*	*Horn*
Aircraft	**71.6**	0	16.1	7.8	4.5
Motorcycle	0	**71.3**	21	0	7.7
Car	23	0	**70.4**	6.2	0.4
Train	21.5	0	0	**78.5**	0
Horn	0	3	16	6	**75**

Table 5
The confusion matrix with respect to the environmental group of the two-stage topology based on the MFCC set.

Presented	Predicted (%)		
	Crowd	*Thunder*	*Wind*
Crowd	**86.3**	11	2.7
Thunder	8.6	**91.4**	0
Wind	24.4	0	**75.6**

4.1. *Enhancement of the feature extraction module*

The following step of the system development phase is focused on the feature set. Our basic set is the MFCC one since it manages to capture a compact picture of the energy spectrum. The additional sets provide information not captured by the MFCC set which may be critical for recognition as well as when the system operates under adverse conditions. To this end, we employed the three groups of parameters, explained in 3.1, simultaneously. The frame was 30 ms with a step of 10 ms, the size of the *FFT*

Please cite this article in press as: S. Ntalampiras, Universal background modeling for acoustic surveillance of urban traffic, Digital Signal Process. (2014), http://dx.doi.org/10.1016/j.dsp.2014.05.003

Table 6
The original dimension of each feature set and the one retaining the 95% of the variance after the application of the DCT.

Feature set	# of original coefficients	# of the retained DCT coefficients
MFCC + ΔMFCC + $\Delta\Delta$MFCC	39	20
MPEG-7 LLDs	27	21
Perceptual Wavelet Packets	136	71
MFCC + ΔMFCC + $\Delta\Delta$MFCC + MPEG-7 LLDs	66	38
MFCC + ΔMFCC + $\Delta\Delta$MFCC + MPEG-7 LLDs + PWP	202	97

Table 7
The confusion matrix with respect to the first phase of the two-stage topology based on UBM and feature fusion (in the parentheses) respectively.

Presented	Predicted (%)		
	Mechanic	*Environmental*	*Crash*
Mechanic	**98** (94)	1.3 (2.3)	0.7 (3.7)
Environmental	1.1 (2.1)	**95.4** (91.5)	3.5 (6.4)
Crash	0 (0)	1.8 (6.7)	**98.2** (93.3)

Table 9
The confusion matrix with respect to the environmental group of the two-stage topology based on UBM and feature fusion (in the parentheses) respectively.

Presented	Predicted (%)		
	Crowd	*Thunder*	*Wind*
Crowd	**94.3** (89.8)	5.2 (8.6)	0.5 (1.6)
Thunder	6 (7)	**94** (93)	0 (0)
Wind	14.7 (16.8)	0 (0)	**85.3** (83.2)

was 512 while the data were windowed according to the Hamming technique. The discrete cosine transform (DCT) was the final processing stage for every feature set. The smallest number of coefficients retaining 95% of the variance was kept. This quantity was measured on the training data and both the original and DCT coefficient numbers are shown in Table 6.

During the enhancement of the feature set we examined the discriminative capabilities of the MFCC set enhanced by the MPEG-7 one. This process provided a significant improvement raising the overall recognition rate to 74.4%. Subsequently we appended the PWP set which offered even higher recognition accuracy (79.15%). The corresponding confusion matrices are shown in Tables 7–9. The results reveal that the pattern which is to be modeled and subsequently classified becomes more evident when a multidomain representation is used. The constructed HMMs are able to approximate the multidomain patterns and subsequently discover them inside novel audio sequences.

4.2. Enhancement of the audio pattern recognition module

The next stage towards the improvement of the surveillance framework investigated the universal background modeling (UBM) technique as applied to the case of generalized sound recognition technology. To this end we initially created an HMM with a large number of states and Gaussian functions by pooling equally distributed data from all sound classes. Subsequently the particular model was adapted according to the needs of the topology designed in the first experimental phase. For example when we needed to construct the mechanic model, data was pooled from the respective categories (aircraft, car, motorcycle, train and horn). Care was taken when acquiring the data in order to obtain quantities balanced among the various classes. This way the constructed model could not favor one or more of them. The parameters of the universal HMM were selected from the following sets: number of states: $\{5, 6, 7, 8, 9, 10\}$ and number of Gaussian components: $\{64, 128, 256, 512, 1024, 2048\}$. The final selection was made based on the highest recognition rate criterion. A descriptive example of the way the UBM components are adapted to fit the distribution of a specific sound class (e.g. class A) is shown in Fig. 2.

The results of the UBM application on the two-stage topology are shown in Tables 7–9. The fused feature set including the MFCCs, the MPEG-7 LLDs and the PWP descriptors was employed. The average recognition rate reached 88.5% which is relatively high considering the number of classes and the intra-class diversity. By observing the confusion matrices, we can infer that the majority of the errors refer to sound categories which, in general, include recordings with similar characteristics, e.g. car-movement and motorcycle-movement sound events. The recognition rate of the first stage is 97.2% which comprises a considerable improvement with respect to the simple modeling technique (92.9%). It is important to note that the UBM modeling scheme offered a performance gain with respect to every classification task. This is due to the fact that a class-specific model which is derived from the UBM outputs low log-likelihoods when it processes data from other sound classes. In other words, if we consider the UBM as covering the class-independent space (a wide range of sound classes), then its adaptation will tune the parameters associated with the adaptation data and favor the target audio class dramatically. On the contrary when another audio class is presented, the non-adapted mixture components will be activated producing much lower probabilities since they are merely copied by the UBM during the adaptation procedure. These do not contribute evidence either toward or against the hypothesized class. In the case the model is trained only on class-specific data, naturally it will output even higher log-likelihoods than the model which is adapted from the UBM one. However this can lead to incorrect values when the unseen data is presented for testing. In addition, during this series of experiments it was shown that when combining parameters of multiple domains, the recognition rate is increased since they have the ability to capture various aspects of the structure of the audio signal. Lastly it is interesting to note that when the UBM scheme was applied alone, it offered superior improvement than the sole application of the fused feature set. More specifically the UBM scheme based on the MFCC set of parameters provided an overall rate of 81.5% while the fused feature combined with class-specific HMMs 79.15% (see Section 4.1). The following sections comment on the design of exhaustive crash detection experiments, the scope of which was to simulate operational conditions.

Table 8
The confusion matrix with respect to the mechanic group of the two-stage topology based on UBM and feature fusion (in the parentheses) respectively.

Presented	Predicted (%)				
	Aircraft	*Motorcycle*	*Car*	*Train*	*Horn*
Aircraft	**91.6** (81.6)	0 (0)	5.1 (8.4)	3.3 (6.6)	0 (4.4)
Motorcycle	0 (0)	**86** (76.7)	11.3 (18.7)	0 (0)	2.7 (4.6)
Car	7.1 (15.6)	0 (0)	**90.9** (80.5)	2 (3.6)	0 (0.3)
Train	9.4 (14.6)	0 (0)	0 (0)	**90.6** (85.4)	0 (0)
Horn	0 (0)	0 (3)	3.4 (9)	0.6 (4)	**96** (84)

Please cite this article in press as: S. Ntalampiras, Universal background modeling for acoustic surveillance of urban traffic, Digital Signal Process. (2014), http://dx.doi.org/10.1016/j.dsp.2014.05.003

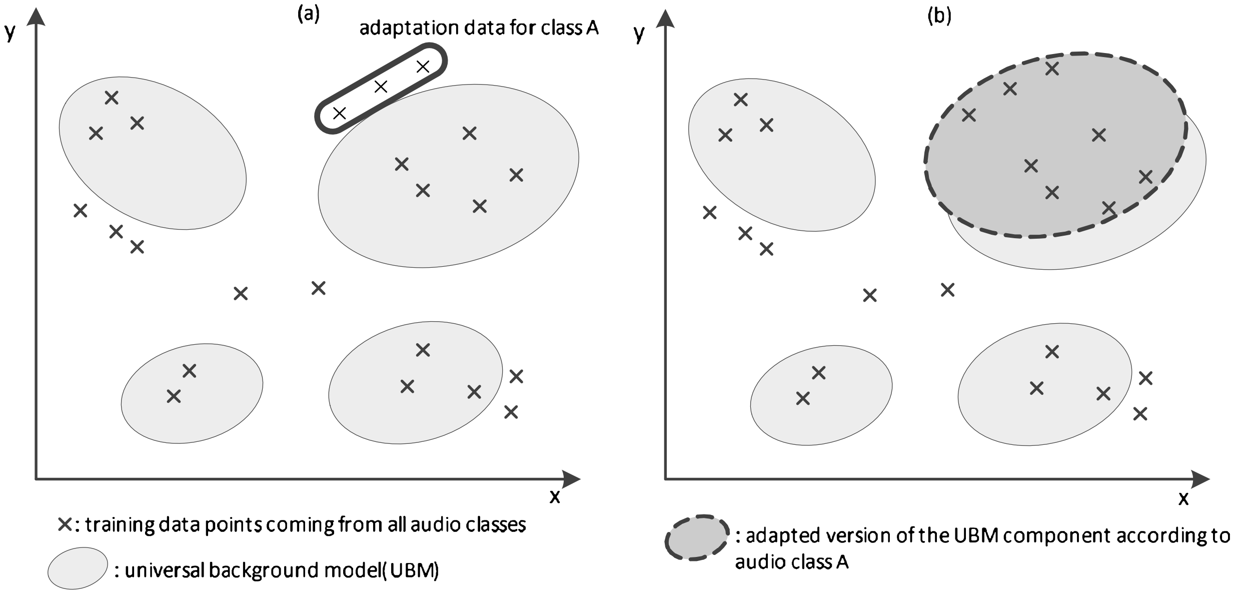
(a)
adaptation data for class A
y
x
×: training data points coming from all audio classes
: universal background model(UBM)
(b)
y
x
: adapted version of the UBM component according to audio class A

Fig. 2. An informative example of the way the UBM is adapted to represent a specific sound category. Subfigure (a) shows the distribution of all training data points (x) and how the universal model approximates them. Subfigure (b) demonstrates the adaptation of the UBM in order to estimate a class-specific model.

5. The case of crash sound events

A problem which falls under the context of the surveillance of urban traffic is the detection of sounds associated with crash events. We intent to apply the methodology described so far onto the specific problem and assess its effectiveness. This section is concentrated on the specific topic, which comprises a useful operation of a traffic surveillance system and may help to deal with a potentially catastrophic event. The basic aims are to limit the consequences of the hazardous situation as much as possible by promptly notifying the authorized personnel and to restore the traffic conditions. In case the system detects a sound event that is related to a crash, a warning message is sent to the authorized personnel for taking the appropriate measures while suggestions may be presented through a decision support interface. As traffic congestion may be caused by accidents, their prompt detection is of significant importance. For example, it can be employed for informing the travelers to avoid a specific route through appropriate electronic signing.

In order to evaluate the ability of the proposed framework to detect crash sound events we simulated situations which include the specific sound events by artificially merging the corresponding signals with the rest of the classes. Our main interest was to observe the response of the system even at adverse conditions, thus we considered different energy ratios during the mixing of the audio signals. More in detail the signal-to-noise ratios (SNR) were 0, 5, and 15 dB. After merging, each output was normalized by its maximum value in order to adjust the overall volume of the specific recording so that the strongest peak was at full level (gain normalization). The feature extraction phase follows and the signal is classified according to the two-stage topology. Its design was based on the statistical models which provided the highest recognition accuracy during the previous experimental phase. Thus the models belonging to the first layer shown in Fig. 1 (mechanic, environmental and crash) were used.

We artificially merged every recording which belongs to the crash sound class with a part of a sound event which belongs to a different class. The part of the signal had the same size to the crash sound event and was chosen in a random way. This process was repeated 100 times for each recording of the crash sound class resulting to $149 \times 100 = 14\,900$ different test instances, which ensured the production of reliable detection results.

One of the most reliable ways to evaluate a system which targets at event detection, independently of the event, is the Detection Error Trade-off (DET) curves [24]. Other evaluation metrics, such as confusion matrices fail to provide a complete picture of the two kinds of errors which are present: a) the kind where a crash sound event is present but it is not detected and b) the kind where a crash sound event is not present, nonetheless it is detected. DET curves comprise an alternative to the Receiver Operator Characteristics (ROC) curve. In the case of ROC curves, the True-Positive-Rate ($R_{TP} = \frac{T_P}{F_N+T_P}$, percentage of the correct classified test cases from all of those which are positive in reality) is plotted in relation to the False-Positive-Rate ($R_{FP} = \frac{F_P}{F_P+T_N}$, percentage of test cases that are negative in reality and wrongly classified as positive by the detector) in dependence of a parameter, typically a threshold T. The decision of the system regarding a possible detection or non-detection is based on the value of T, thus its optimal value should be found. DET curves try to present the trade-off between missed detections and false alarms and the operational point is located where the average of the missed detection and false alarm rates is minimized. In other words this average corresponds to the cost function of a DET curve. However a DET curve is meaningful when a large number of target events (crash sound events) is available along with an almost equal quantity of non-target events. Then, the performance is depicted accurately. A convenient property of the DET curves is that unlike the ROC ones they are linear or close to linear. This fact is beneficial as it allows for an easy and safe choice of the system/configuration with the best detection ability.

For the needs of this phase of the experiments, target (crash) and non-target (mechanic and environmental) events were provided as input to the two-stage probabilistic framework. Based on the log-likelihoods which are produced by the models of the first-stage of the topology, we designed the DET curves with respect to three noise levels. The associated SNR values were 0, 5 and 15 dB and the DET curves are plotted on the same diagram (see Fig. 3). As we can see the higher the SNR value, the higher the reliability of the detector. Due to the linear nature of the DET curves, this fact is easily observable. More in detail, the Equal Error Rates (EER) are 6.6%, 2.2% and 0.9% for SNR values 0, 5, 15 dB respectively. The miss probability and the false alarm probability are kept in

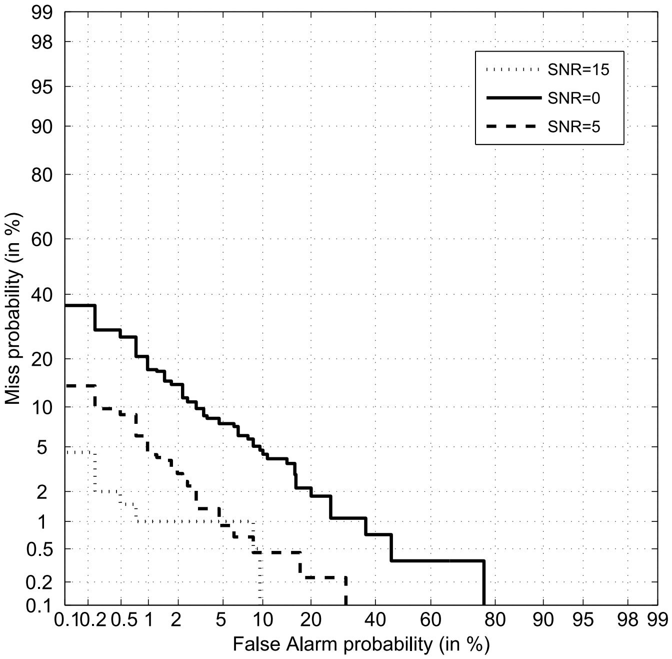

Fig. 3. Illustration of the incident detection capabilities of the two-stage framework. Crash sound events are merged with the rest of the audio classes at three SNR levels (0, 5 and 15 dB). The target class is crash while the non-target classes are mechanic and environmental. The respective EERs are 6.6%, 2.2% and 0.9%.

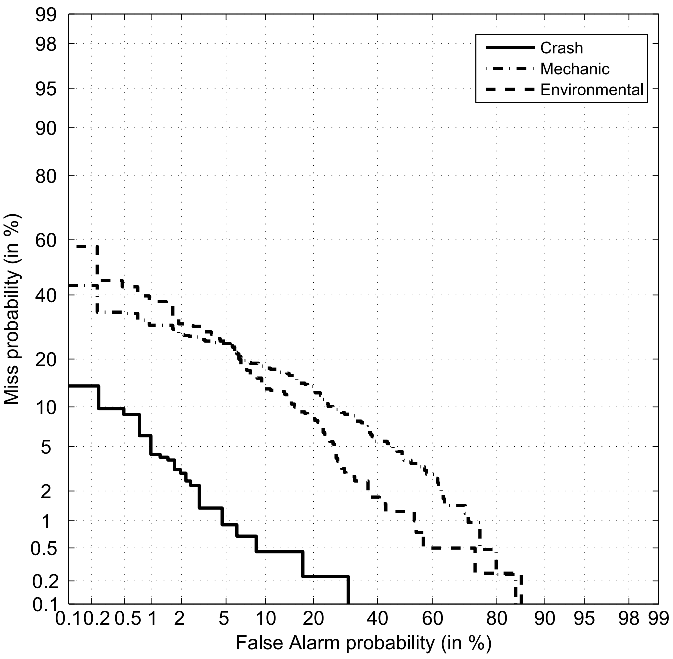

Fig. 4. Illustration of the detection capabilities of the first classification stage. The following classes were used as the target ones: crash, mechanic and environmental. These are merged with the rest of the audio classes at $SNR = 5$ dB. The respective EERs are 2.3%, 15.2%, 11.5%.

quite good levels even at highly noisy conditions ($SNR = 0$ dB). The recordings which are a result of merging conducted at $SNR = 5$ dB represent the real-world conditions quite well (to the extent that can be assessed by a human listener). At the particular ratio, our system provided a relatively low EER that shows reliable detection of the sound events of interest. To conclude with the detection of isolated crash sound events, we infer that the results are quite promising and reveal the merits of the two-stage classification topology, in which features with diverse characteristics are incorporated.

6. Application of the surveillance framework on continuous audio flows

This section describes how the proposed surveillance framework is applied onto audio streams of potentially unlimited duration. Our purpose is to establish a generic methodology that not only elaborates on isolated sound events but also on continuous audio flow. With this experiment we shift our focus from the classification of isolated sound events (which is usually encountered in the audio recognition literature) to the simulation of real-world conditions to some extent. We wish to examine the detection of classes which are in line with the purpose of such a framework, i.e. *mechanic*, *environmental*, *crash* as well as *aircraft*, *car*, *motorcycle* and *train* sound events.

We are concentrated on the simulation of real-world conditions to some extent. Toward this end we concatenated sound events belonging to all the sound categories, resulting at 50 sequences of approximately 15 minutes duration each. Each sound event of interest was merged (including gain normalization) with one belonging to the rest of the classes in the way which was described in Section 5 while the SNR was set at 5 dB. The 50 audiostreams contained various changes between target and non-target sound events, subsequent target events as well as subsequent non-target events. For the detection of the target sound events we used a method which is usually called *detection-by-classification* [4], i.e. each frame was fed as input to the system for automatic processing and eventual recognition. This type of processing can deal with the two fold problem which is presented when a monitoring system needs to process sequences of unlimited duration: both the duration and start time of a target sound event are not a-priori known. Thus it is of significant difficulty to detect urban environmental sound events since their duration varies greatly. While the bottom layer of the system processes each frame, the next layer concatenates decisions with the same classification tag and creates a list which includes the recognized events. For each event on the list, we assign a confidence metric which may be useful to the personnel operating the surveillance machine. This metric is equal to the sum of the log-likelihoods associated to the specific events normalized by the number of frames. It is worth mentioning that isolated frame decisions are thought to comprise outliers and are automatically discarded (30 ms do not comprise a logical duration for any event).

The DET curves associated with the first stage of the proposed framework are shown in Fig. 4 and the ones associated with the mechanic sound classes are shown in Fig. 5. More precisely the EERs for classes mechanic, environmental and crash are 15.2%, 11.5%, 2.3% and for classes aircraft, motorcycle, car, train are 2.4%, 7.5%, 0.7%, 2.2%. The achieved EERs are quite low, an observation demonstrating the efficacy of the proposed surveillance framework. We conclude that the presented results are more than satisfactory, underline the importance of the probabilistic recognition topology and indicate that the system provides an excellent basis for a practical implementation (e.g. an embedded system) able to operate under real-world conditions.

7. Conclusions

We believe that robust automatic traffic surveillance systems have the potential to boost the current traffic management frameworks. Large-scale developments (e.g. microphones on traffic lights/signs, on surveillance cameras, etc.) can have a heavy impact on ITS applications. The acoustic modality is able to offer valuable information regarding the ongoing situations and can be used in parallel with the visual one. It has numerous applications covering various needs (from parking guidance to making available

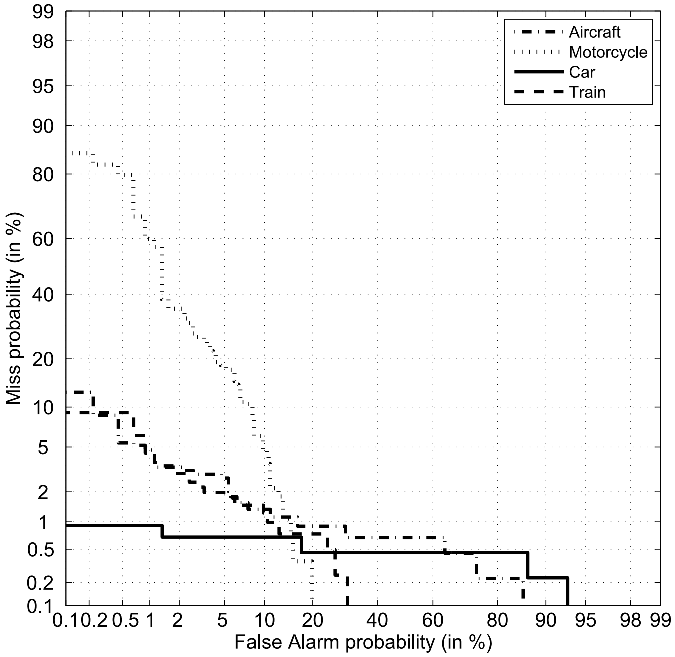

Fig. 5. Illustration of the detection capabilities of the two-stage framework with respect to the mechanic sound events: aircraft, motorcycle, car and train sound events. These are merged with the rest of the audio classes at *SNR* = 5 dB. The respective EERs are 2.4%, 7.5%, 0.7%, 2.2%.

traveling information including weather phenomena and detecting accidents).

We analyzed a novel methodology which achieves acoustic surveillance of urban traffic capable of addressing several issues, such as: detection and identification of a gamut of vehicle types, as well as detection of crash sound events. It allows for enumerating and indexing these vehicles types, it does not suffer from problems related to the vision-based surveillance systems (e.g. occlusion, lighting condition, out of camera's angle of view, etc.) while it is able to operate under a wide range of highly noisy conditions. On the contrary when the objectives include the exact measurements of vehicle flow, velocity and density, the audio modality can only play a complementary role. The analysis of urban traffic surveillance in the long term could provide valuable information for a deep understanding of current urban traffic networks and assist future town planning.

Obviously, additional sound classes may exist in a real environment. Nonetheless the proposed framework is flexible since new categories can be easily incorporated as long as an adequate amount of training data is available to create the respective statistical models.

Intelligent monitoring systems are of significant importance, especially in the context of smart cities [9]. Further exploitation of the generalized sound recognition technology in the specific research field is feasible as it has not yet reached its full potential. Application of signal separation and source tracking [38] comprise part of our future works. We also intent to experiment on designing audio signatures for vehicles of interest (e.g. ambulances, police vehicles or vehicles associated with illegal activities) towards their reliable detection inside the urban road plan.

References

[1] A. Averbuch, V.A. Zheludev, N. Rabin, A. Schclar, Wavelet-based acoustic detection of moving vehicles, Multidimens. Syst. Signal Process. 20 (1) (Mar. 2009) 55–80, http://dx.doi.org/10.1007/s11045-008-0058-z.

[2] D. Botteldooren, L. Dekoninck, D. Gillis, The influence of traffic noise on appreciation of the living quality of a neighborhood, Int. J. Environ. Res. Public Health 8 (3) (2011) 777–798, http://www.mdpi.com/1660-4601/8/3/777.

[3] M. Bramberger, A. Doblander, A. Maier, B. Rinner, H. Schwabach, Distributed embedded smart cameras for surveillance applications, Computer 39 (2) (Feb. 2006) 68–75.

[4] T. Butko, C. Nadeu, Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion, EURASIP J. Audio Speech Music Process. 2011 (2011) 1, http://dblp.uni-trier.de/db/journals/ejasmp/ejasmp2011.html#ButkoN11.

[5] M. Casey, General sound classification and similarity in MPEG-7, Organ. Sound 6 (2) (Aug. 2001) 153–164, http://dx.doi.org/10.1017/S1355771801002126.

[6] M. Casey, MPEG-7 sound-recognition tools, IEEE Trans. Circuits Syst. Video Technol. 11 (6) (2001) 737–747.

[7] S. Chen, Z. Sun, B. Bridge, Traffic monitoring using digital sound field mapping, IEEE Trans. Veh. Technol. 50 (6) (Nov. 2001) 1582–1589.

[8] P. Chordia, Segmentation and Recognition of Tabla Strokes, University of London, London, UK, Sept. 2005, pp. 107–114, http://ismir2005.ismir.net/proceedings/1137.pdf.

[9] H. Chourabi, T. Nam, S. Walker, J.R. Gil-García, S. Mellouli, K. Nahon, T.A. Pardo, H.J. Scholl, Understanding smart cities: an integrative framework, in: HICSS, IEEE Computer Society, 2012, pp. 2289–2297, http://dblp.uni-trier.de/db/conf/hicss/hicss2012.html#ChourabiNWGMNPS12, 2012.

[10] B. Coifman, D. Beymer, P. McLauchlan, J. Malik, A real-time computer vision system for vehicle tracking and traffic surveillance, Transp. Res., Part C, Emerg. Technol. 6 (4) (1998) 271–288, http://www.sciencedirect.com/science/article/pii/S0968090X98000199.

[11] R. Cucchiara, M. Piccardi, P. Mello, Image analysis and rule-based reasoning for a traffic monitoring system, IEEE Trans. Intell. Transp. Syst. 1 (2) (2000) 119–130.

[12] Databases, Sound ideas, http://www.sound-ideas.com, 2012.

[13] K.B. Eom, Analysis of acoustic signatures from moving vehicles using time-varying autoregressive models, Multidimens. Syst. Signal Process. 10 (1999) 357–378, http://dx.doi.org/10.1023/A:1008475713345.

[14] J.-L. Gauvain, C.-H. Lee, Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Trans. Speech Audio Process. 2 (2) (1994) 291–298, http://dblp.uni-trier.de/db/journals/taslp/taslp2.html#GauvainL94.

[15] M. Ghiurcau, C. Rusu, Vehicle sound classification. Application and low pass filtering influence, in: International Symposium on Signals, Circuits and Systems, ISSCS 2009, July 2009, pp. 1–4.

[16] P. Hills, P. Blythe, For whom the road tolls? The future for road congestion charging in the UK, Igenia 14 (2012) 21–28.

[17] K. Hu, D. Wang, Incorporating spectral subtraction and noise type for unvoiced speech segregation, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, 2009, pp. 4425–4428.

[18] K. Hu, D. Wang, Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction, IEEE Trans. Audio Speech Lang. Process. 19 (6) (Aug. 2011) 1600–1609.

[19] T. Isobe, J. ichi Takahashi, Text-independent speaker verification using virtual speaker based cohort normalization, in: EUROSPEECH, ISCA, 1999, http://dblp.uni-trier.de/db/conf/interspeech/eurospeech1999.html#IsobeT99.

[20] S. Kamijo, Y. Matsushita, K. Ikeuchi, M. Sakauchi, Traffic monitoring and accident detection at intersections, in: Proceedings, IEEE/IEEJ/JSAI International Conference on Intelligent Transportation Systems, 1999, pp. 703–708.

[21] H.-G. Kim, T. Sikora, Comparison of MPEG-7 audio spectrum projection features and MFCC applied to speaker recognition, sound classification and audio segmentation, in: Proceedings, vol. 5, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04), 2004, V-925-8.

[22] A. Klausner, A. Tengg, C. Leistner, S. Erb, B. Rinner, An audio-visual sensor fusion approach for feature based vehicle identification, in: IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007, Sept. 2007, pp. 111–116.

[23] F. Kraft, R. Malkin, T. Schaaf, A. Waibel, Temporal ICA for classification of acoustic events in a kitchen environment, in: INTERSPEECH 2005 – Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4–8, 2005, ISCA, 2005, pp. 2689–2692.

[24] A.F. Martin, G.R. Doddington, T. Kamm, M. Ordowski, M.A. Przybocki, The DET curve in assessment of detection task performance, in: G. Kokkinakis, N. Fakotakis, E. Dermatas (Eds.), EUROSPEECH, ISCA, 1997, http://dblp.uni-trier.de/db/conf/interspeech/eurospeech1997.html#MartinDKOP97.

[25] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget, J. Cernocky, Full-covariance UBM and heavy-tailed PLDA in i-Vector speaker verification, in: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2011, pp. 4828–4831.

[26] P. Mohan, V.N. Padmanabhan, R. Ramjee, Nericell: rich monitoring of road and traffic conditions using mobile smartphones, in: Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems, SenSys '08, ACM, New York, NY, USA, 2008, pp. 323–336, http://doi.acm.org/10.1145/1460412.1460444.

[27] M.E. Munich, Bayesian subspace methods for acoustic signature recognition of vehicles, in: Proc. of the 12th European Signal Processing Conf. (EUSIPCO 2004), Vienna, Austria, Sept. 6–10, 2004.

[28] B.F. Necioglu, C.T. Christou, E.B. George, G.M. Jacyna, Vehicle acoustic classification in netted sensor systems using Gaussian mixture models, in: I. Kadar

(Ed.), Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 5809, May 2005, pp. 409–419.

[29] S. Ntalampiras, I. Potamitis, N. Fakotakis, Automatic recognition of urban environmental sound events, in: International Association for Pattern Recognition Workshop on Cognitive Information Processing, Santorini, Greece, June 9–10, 2008, pp. 110–113.

[30] S. Ntalampiras, I. Potamitis, N. Fakotakis, Exploiting temporal feature integration for generalized sound recognition, EURASIP J. Adv. Signal Process. 2009 (2009).

[31] M.T. Obaidat, Spatial mapping of traffic noise levels in urban areas, J. Transp. Res. Forum 47 (2) (2008) 89–102.

[32] I. Potamitis, T. Ganchev, Generalized recognition of sound events: approaches and applications, pp. 41–79, http://dx.doi.org/10.1007/978-3-540-78502-6_3, 2008.

[33] O. Report, Highway statistics 2000, FHWA report, 2000, http://www.fhwa.dot.gov/ohim/hs00/index.htm.

[34] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, in: Digital Signal Processing, 2000, p. 2000.

[35] A. Rosenberg, S. Parthasarathy, Speaker background models for connected digit password speaker verification, in: Conference Proceedings, vol. 1, 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-96, May 1996, pp. 81–84.

[36] C. Torrence, G.P. Compo, A practical guide to wavelet analysis, Bull. Am. Meteorol. Soc. 79 (1) (Jan. 1998) 61–78, http://dx.doi.org/10.1175/1520-0477(1998)079%3C0061:APGTWA%3E2.0.CO;2.

[37] V. Tyagi, S. Kalyanaraman, R. Krishnapuram, Vehicular traffic density state estimation based on cumulative road acoustics, IEEE Trans. Intell. Transp. Syst. 13 (3) (2012) 1156–1166.

[38] Y. Wang, K. Han, D. Wang, Exploring monaural features for classification-based speech segregation, IEEE Trans. Audio Speech Lang. Process. 21 (2) (2013) 270–279.

[39] H. Wu, J.M. Mendel, Classification of battlefield ground vehicles using acoustic features and fuzzy logic rule-based classifiers, IEEE Trans. Fuzzy Syst. 15 (1) (Feb. 2007) 56–72.

[40] H. Wu, M. Siegel, P. Khosla, Vehicle sound signature recognition by frequency vector principal component analysis, IEEE Trans. Instrum. Meas. 48 (5) (Oct. 1999) 1005–1009.

[41] S.J. Young, G. Evermann, M.J.F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P.C. Woodland, The HTK Book, Version 3.4, Cambridge University Engineering Department, Cambridge, UK, 2006.

Stavros Ntalampiras received the Engineer's and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Patras, Patras, Greece, in 2006 and 2010, respectively. He has been a post-doc researcher with the System Architectures Group, Department of Electronics, Information and Bioengineering of the Politecnico di Milano. Currently, he is a post-doc researcher working at the Joint Research Center of the European Commission. He has authored more than 35 articles in peer-reviewed international journals and conference proceedings. His research interests include content-based signal processing, computational intelligence, fault diagnosis, and computer audition.

Please cite this article in press as: S. Ntalampiras, Universal background modeling for acoustic surveillance of urban traffic, Digital Signal Process. (2014), http://dx.doi.org/10.1016/j.dsp.2014.05.003