# AUTOMATIC RECOGNITION OF URBAN ENVIRONMENTAL SOUNDS EVENTS

*Stavros Ntalampiras*[†], *Ilyas Potamitis*[±], *Nikos Fakotakis*[†]

[†]Wire Communicatioons Laboratory, University Of Patras, {dallas, fakotaki}@wcl.ee.upatras.gr,
[±]Department of Music Technology and Acoustics, Technological Educational Institute of Crete, potamitis@stef.teicrete.gr

## ABSTRACT

Computer audition is an evolving and relatively new research field with many new applications. It would be of great convenience to live in an environment that can change automatically based on its "auditory sense". In this work we propose a novel framework for automatic recognition of urban soundscenes. Our system facilitates a hierarchical classification schema while the performance of two well known feature sets is compared. A new post- processing algorithm to enhance the discrimination quality of MPEG-7 features is proposed and shown to provide improved results. Our approach is examined utilizing a compact testing procedure while MPEG-7 LLDs reach higher recognition rates than MFCCs.

***Index Terms***— Computer Audition, Environmental sound recognition, MFCC, MPEG-7, Hidden Markov Models (HMM)

## 1. INTRODUCTION

Nowadays we experience a lot of different types of urban sounds in our everyday life (car, motorcycle, crowd etc). Humans can effectively differentiate them quite effortless utilizing only the auditory sense. Think as a paradigm the situation were one is waiting at a traffic light. Using incoming sounds alone one is able to understand that a car is passing and a dog is barking in the presence of a horn sound. The general scope of our work is to build up a system that has the ability to automatically "understand" its surrounding environment by taking under consideration the sounds it "hears" alone. The area of computer audition faces an increasing demand in numerous applications (robotic awareness, environmental monitoring, media annotation etc) thus becoming a research field of great importance.

Over the past decades a great deal of work has been published in the area of content-based audio classification. Eronen et al [1] explore an audio based recognition system used for classification of 24 urban contexts. They utilize several simplistic low-dimensional features as well as standard spectral descriptors along with an HMM-based classification scheme achieving 58% recognition rate. A framework for frame-level classification of noises belonging to five categories is presented in [2]. Line Spectral Frequencies (LSFs) in combination with a decision tree classifier were employed resulting in 88.1% classification accuracy. A method based on three MPEG-7 audio low-level descriptors (spectrum centroid, spectrum spread and spectrum flatness) is presented in [3]. For classification scheme a fusion of support vector machines and *k* nearest neighbour rule is adopted in order to assign a specific sound into predefined classes of common kinds of home environmental sounds.

In this work we employed MFCCs and MPEG-7 descriptors. Our aim is to identify which feature set contains more discriminative information to serve the task of recognition of urban soundscenes. The rest of the paper is organized as follows. The next three sections describe the overall architecture of our implementation, the feature extraction methodology and the recognition procedure. Detailed analysis of the evaluation method as well as the results is given in the last part of the paper.

## 2. SYSTEM OVERVIEW

In this section the architecture of our system which makes possible automatic recognition of urban soundscenes is described. We approach the issue based on the way that humans categorize subconsciously their surrounding environment using *only* the perceived acoustic information. Our goal is to distinguish scenes belonging to eight different classes: aircraft, motorcycle, car, crowd, thunder, wind, train, horn.

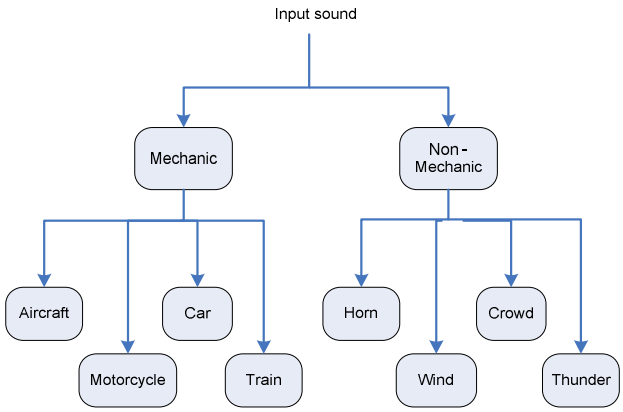


Figure 1. Tree of Categories

We utilize a hierarchical classification scheme consisting of two stages and derived from a perceptual point of view of

the classes (Fig. 1). The first stage classifies sounds into two categories (mechanic and non-mechanic) while the second one completes the rest of the classification process producing the leaf-class. Furthermore our implementation includes preprocessing of the audio signals, feature extraction, elimination of silence frames, principal component analysis (PCA) and training different kinds of classifiers whose parameters represent the *a-priori* knowledge we have available for the different audio classes.

## 3. FEATURE EXTRACTION AND SILENCE ELIMINATION

In order to evaluate the performance of the selected feature sets in the task of environmental sound recognition the same preprocessing method must be applied. Thus all the parameters were kept the same for both feature extraction processes. Signals are cut into frames of 30ms with 10ms time shift between two successive frames while they are hamming windowed following MPEG-7 standard recommendations.

Silence is considered to be "noise" in this particular task making harder the process of modeling, hence reducing the probability of correct classification. Subsequently a simple amplitude-based silence detection algorithm is used applied onto each sample. If all sample's amplitudes of a specific frame are bellow 4% of the average signal's amplitude, the frame is considered to be silent thus not involved in training nor testing procedure. On top of that a standard version of PCA is applied onto the MPEG-7 feature vector for the purpose of dimensionality reduction.

### 3.1. MFCCs

This feature set is composed of the first twelve Mel frequency cepstral coefficients plus frame's total energy. For MFCC's derivation we compute the power of the Short time Fourier transform (STFT) for every frame and pass them through a triangular Mel scale filterbank so that signal components which play an important role to human perception are emphasized. Subsequently, the log operator is applied and we exploit the energy compaction properties that Discrete Cosine transform (DCT) benefits in order to decorrelate and represent the majority of the frame-energy with just a few of its coefficients. Lastly the most important twelve coefficients are kept and in combination with frame's energy a thirteen-dimension vector is formed.

### 3.2. MPEG-7 feature set

The main idea behind MPEG-7 standard is the creation of a method for automatic audio content description capable of providing solutions in numerous problems such as indexing, retrieval and classification. In this work we take advantage of the following descriptors (Low Level Descriptors – LLDs):

- Audio Waveform (AWF)

This constitutes compact description of the shape of an audio signal by computing the minimum and maximum samples within successive non-overlapping frames.

- Audio Power (AP)

It is a temporal descriptor representing the evolution of the signal's sampled data during time and is computed by the following formula:

$$AP(l) = \frac{1}{N_{hop}} \sum_{n=0}^{N_{hop}-1} |s(n + lN_{hop})|^2 \quad (0 \le l \le L\text{-}1), \qquad (1)$$

where $L$ is the total number of frames and $N_{hop}$ the number of time samples between two successive frames

- Audio Spectrum Envelope (ASE)

This series of features belong to the basic spectral descriptors and is derived for the generation of a reduced spectrogram of the original audio signal. It is a log-frequency power spectrum and calculated by summing the energy of the original power spectrum within a series of logarithmically distributed frequency bands utilizing a predefined resolution.

- Audio Spectrum Centroid (ASC)

The center of the log-frequency spectrum's gravity is given by this descriptor. Omitting power coefficients bellow 62.5Hz (which are represented by a single coefficient) makes able the avoidance of the effect of a non-zero DC component. For a given frame the ASC is defined from the modified power coefficients and their frequencies as:

$$ASC = \sum_i \log_2(f_i/1000)p_i) \Big/ \sum_i p_i \qquad (2)$$

where $p_i$ is the power spectrum while $f_i$ represent the corresponding frequencies.

- Audio Spectrum Spread (ASS)

ASS or instantaneous bandwidth is a measure of signal's spectral shape and corresponds to the second central moment of the log-frequency spectrum. It is computed by taking the root mean square (RMS) deviation of the spectrum from its Centroid:

$$ASS = \sqrt{\sum_i ((\log_2(f_i/1000) - ASC)^2 p_i \Big/ \sum_i p_i} \qquad (3)$$

- Audio Spectrum Flatness (ASF)

This descriptor is a measure of how flat a particular portion of the signal is and represents the deviation of the signal's power spectrum from a flat shape. The power coefficients are taken from non-overlapping frames while the spectrum is divided into ¼-octave resolution logarithmically spaced overlapping frequency bands. The ASF is derived as the ratio of the geometric mean and the arithmetic mean of the spectral power coefficients within a band.

$$ASF = \sqrt[N]{\prod_{n=1}^{N} C_n} \Big/ \frac{1}{N} \sum_{n=1}^{N} C_n \qquad (4)$$

where $N$ is the number of coefficients within a subband and $c_n$ is the $n$-th spectral power coefficient of the subband. This

feature can efficiently differentiate between noise (or impulse) and harmonic sounds and we should take into account that a large deviation from a flat shape generally depicts *tonal* sounds.

The next two descriptors reflect upon the harmonic structure of periodic sounds and can efficiently differentiate between *harmonic* (music, voiced speech) and *non-harmonic* (noise, unvoiced speech) sounds:

- Harmonic Ratio (HR)

This corresponds to the proportion of harmonic components in the power spectrum. Its extraction for every frame is standardized in the following way: the maximum value of the normalized autocorrelation function is computed overall the specific frame. If the signal is purely periodic its peak values will be at lags $m$ (which denotes the index of autocorrelation) corresponding to multiples of fundamental period $T_0$. HR will be close to one for harmonic signals and zero for white noise.

- Upper Limit of Harmonicity (ULH)

This feature provides a measure of the frequency value beyond which the spectrum no longer has any harmonic structure. A time domain comb filter (Moorer, 1974) which is tuned to the fundamental period of the signal (taken from the previous descriptor) is utilized and the proportion of its output/input power forms the basis of the computation of ULH.

- Audio Fundamental Frequency (AFF)

For a given and assumed to be periodic portion of the signal AFF consists of an estimation of the fundamental frequency $f_0$. It can be used as an approximation of the pitch of musical sounds and voiced speech.

At this point we propose a post-processing methodology that serves the enhancement of the discriminative ability of the MPEG-7 feature set. The extraction of the MPEG-7 descriptors is followed by log operation as well as the Discrete Cosine transform (DCT) in order to obtain efficient representation of the signal's energy. A common technique for finding patterns in data of high dimensions, principal component analysis (PCA) is utilized on the LLDs. We reduce the dimensions down to thirteen by calculating the projection of the data onto a lower dimensions space created by its most significant eigenvectors. With this procedure the data are transformed to a new coordination system based on the relationships between them. Finally it should be noted that normalization techniques are applied on both feature sets including mean removal and variance scaling.

## 4. CLASSIFICATION SCHEMAS

The first step in the recognition process is based on Gaussian mixture models (GMM) created using a standard version of Expectation Maximization (EM) algorithm with $k$-means initialization. They approximate a probability density function under the assumption that every distribution can be modeled when enough Gaussian distributions are combined. The result is of the form of a weighted sum of $M$ simpler Gaussian densities (components):

$$p(x_t) = \sum_{m=1}^{m=M} \pi_m N(x_t, \mu_m, \Sigma_m) \qquad (5)$$

where $x_t$ is the feature vector at time $t$, $N$ is a Gaussian pdf with mean $\mu_m$, covariance matrix $\Sigma_m$ and $\pi_m$ is the component prior probability. We used eight Gaussian components with 50 iterations of the EM algorithm with $k$-means initialization. In Fig. 2 we can see the next step of the proposed methodology which consists of HMMs created for each class. Each HMM consists of two states combined in a

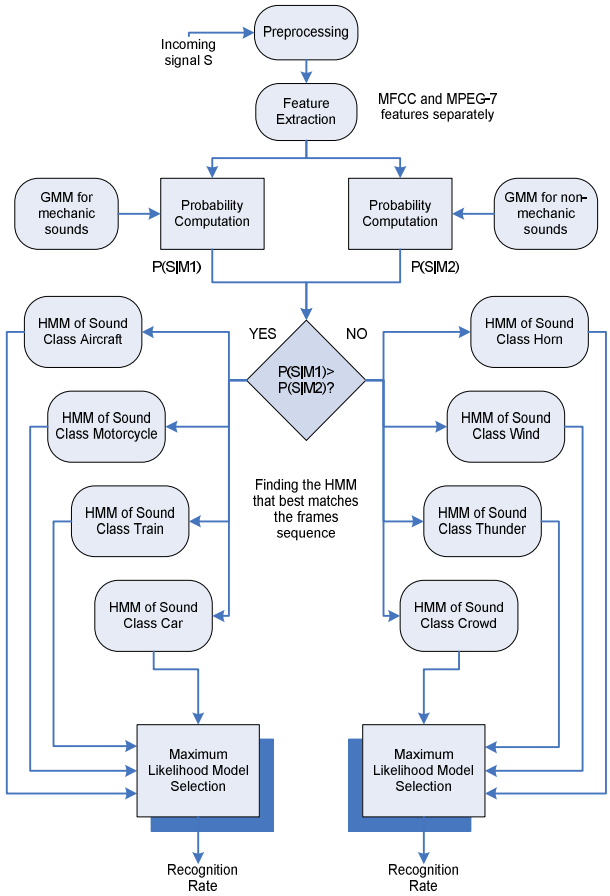


Figure 2. Overall Architecture of the proposed methodology

left-right topology and modeled using sixteen Gaussian modes. HMMs are based on the assumption that the process we are trying to model can be divided into a finite number of states and that their succession over time can be predicted. The training is done with 50 iterations of the Baum-Welch algorithm. The HMM's log-likelihood output tells us how possible is that the order of these specific states generated the input sequence. GMMs and HMMs are created based on P. Baggenstoss' implementation provided at http://www.npt.nuwc.navy.mil/Csf/.

## 5. EXPERIMENTAL SET-UP AND CONCLUSIONS

To test the performance of the proposed structure we collected data from various sources including the BBC Sound Effects Library and recordings found on the internet. Eight categories were organized containing files of 16 kHz and 16 bit analysis while their average length was 25.6 seconds. The classes are aircraft (110), motorcycle (79), car (81), crowd (60), thunder (60), wind (66), train (82) and horns (194), each one includes audio samples with a great variation between them representing real life soundscenes.

| Responded | Aircraft | Motorcycle | Car | Train |
|---|---|---|---|---|
| Aircraft | **57.5** | 17.1 | 17 | 25.4 |
| Motorcycle | 0 | **67.5** | 8.3 | 24.2 |
| Car | 0 | 0 | **57.7** | 42.3 |
| Train | 14.3 | 0 | 0 | **85.7** |
| **Responded** | **Wind** | **Thunder** | **Crowd** | **Horn** |
| Wind | **63.5** | 0 | 36.5 | 0 |
| Thunder | 16.5 | **83.5** | 0 | 0 |
| Crowd | 0 | 0 | **100** | 0 |
| Horn | 0 | 0 | 33 | **66** |

Table 1: Confusion matrices (%) - MFCC

| Responded | Aircraft | Motorcycle | Car | Train |
|---|---|---|---|---|
| Aircraft | **71.6** | 0 | 20.1 | 8.3 |
| Motorcycle | 0 | **71.3** | 21 | 7.7 |
| Car | 33 | 0 | **60.4** | 6.6 |
| Train | 52 | 0 | 0 | **48** |
| **Responded** | **Wind** | **Thunder** | **Crowd** | **Horn** |
| Wind | **89** | 11 | 0 | 0 |
| Thunder | 10 | **90** | 0 | 0 |
| Crowd | 24 | 0 | **76** | 0 |
| Horn | 19 | 0 | 10.3 | **70.7** |

Table 2: Confusion matrices (%) – MPEG-7

During the testing phase incoming sound frames are processed and classified the way we depict in Fig.2. In order to obtain reliable results ten-fold cross validation is employed for all tasks. MPEG-7 descriptors reach 75.3% recognition rate while MFCCs achieved 64.1% for the classification's first stage. The next two Tables show the recognition accuracies for both features sets referring to the second discrimination task. Overall accuracy for the mechanic classes is 67.1% and 62.9% and for the non-mechanic is 78.25% and 81.4% regarding to MFCC and MPEG-7 respectively. It is obvious that both sets tend to confuse mechanic classes and especially car category. In general we observe that both feature sets experience the same classification problems. The best rate for MFCCs and MPEG-7 LLDs is achieved in the crowd and thunder category accordingly. We conclude that post-processing of the MPEG- 7 features improves their discrimination quality and makes them outperform MFCCs in the task of urban soundscene recognition.

Our future work includes separation of overlapping signals, further incorporation of sound classes and exploration of a possible combination of the two feature sets to exploit the most discriminative information they provide.

## 6. REFERENCES


[1] A.J. Eronen, V.T Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, G.Lorho, J. Huopaniemi , "Audio-based context recognition," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 14, no. 1, pp. 321 – 329, January 2006

[2] K. El-Maleh, A. Samouelian, P. Kabal, "Frame level noise classification in mobile environments," Proceedings of the Acoustics, Speech, and Signal Processing, vol. 1, pp. 237-240, 1999

[3] Jia-Ching Wang, Jhing-Fa Wang, Wai-He Kuok and Cheng-Shu Hsu, "Environmental Sound Classification Using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor," International Joint Conference on Neural Networks, 2006.

[4] M. A. Casey, "MPEG-7 sound recognition tools," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 737–747, 2001.

[5] Hyoung-Gook Kim, Nicolas Moreau, Thomas Sikora, "*MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*", Wiley, 2005.